IntechOpen

Products and Services; from R&D to Final Solutions

Edited by Igor Fuerstner

Products and Services; from R&D to Final Solutions

edited by
Igor Fürstner

Products and Services; from R&D to Final Solutions
http://dx.doi.org/10.5772/297
Edited by Igor Fuerstner

Contributors

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in Croatia, 2010 by INTECH d.o.o.
eBook (PDF) Published by IN TECH d.o.o.
Place and year of publication of eBook (PDF): Rijeka, 2019.
IntechOpen is the global imprint of IN TECH d.o.o.
Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from orders@intechopen.com

Products and Services; from R&D to Final Solutions
Edited by Igor Fuerstner

p. cm.

ISBN 978-953-307-211-1

eBook (PDF) ISBN 978-953-51-5144-9

Contents

Preface

The world economy of today is more integrated and interdependent than ever before. The fact that in many industries historically distinct and separate markets are merging into one global market leads towards an environment that offers more opportunities, but is also more complex and competitive than it used to be.

One of the main factors that drive today's economy is technology. If technology is de ned as a practical application of knowledge and the aim is to become really competitive on the global market, there is a need for something more, thus a cutting edge practical application of knowledge would be necessary what the most advanced technology currently available is - high tech.

If the classi cation of high-tech sectors is taken into consideration, it can be noticed that the research activity takes place not only in the so-called high-tech societies such as the United States, Japan, Germany, etc., but also in other regions.

This book is the result of widespread research and development activity, covering different elds of science.

Chapters one to four offer an overview of the research results, covering the aspects of research and development activities in general. They deal with the necessary infrastructure and technologies for distributed knowledge acquisition, factors that determine the success or failure in NPD, methodologies that operationalize sustainable development, approaches of modelling the core factors which in uence the operational competitiveness performance, i.e. manufacturing strategy and transformational leadership with technology level, etc.

Chapters ve to nine discuss several approaches regarding the issues of production systems and product development, procedures for concurrent product and processes development, integration and comparing principles and characteristics of six sigma with Lean Manufacturing, Total Quality Management, Human Resources, Supply Chain Management, Inventory Issues, etc.

Chapters ten to twenty-one discuss various aspects of the practical application of knowledge such as:

- Mass customization
- Nanotechnology
- Environmental protection
- Drilling fluid technology
- Raw material extraction technology
- Virtual reality

- Spoken dialog system
- Text to speech system
- System architecture

October 20, 2010

Editor

Igor Fürstner
Subotica Tech – College of Applied Sciences
Subotica, Serbia

1

Large Scale Distributed Knowledge Infrastructures

Wojtek Sylwestrzak
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland

1. Introduction

Many tend to believe that the current process of development of science is one of the best established and most traditional research procedures that has evolved over years to reach its current almost ultimate perfection. Well, think again. In recent years we are witnessing the beginning of a paradigm shift in the scientific research conduct process. The traditional one-man-show way of performing research, most common in the first half of the 20th century is slowly beginning to disappear in the more advanced disciplines. Instead, modern research is increasingly based on the concepts of massively distributed collaboration, resource sharing, open access to knowledge and achieves new qualities through, often interdisciplinary, compilation research and data reuse. The conversion is becoming possible due to the rapid Internet technologies development of the last decades, electronic communication proliferation but also due to the subtle and indirect influence of the open access culture, originating from the open source software development but popularised through the Internet, and the resulting transformation in the human perception of notions of progress and scholarly work. The European Commission recognizes the importance of open access in research and, in anticipation of the open access mandate for its future programmes, establishes an OpenAIRE open repositories structure to house its future funded research output. One of the key prerequisites to the eventual success of the transformation process is a broad access to knowledge and, consequently, the existence of adequate tools and technologies making this access possible and effective. The availability of such technologies, however, is still lagging behind.

Due to the diversity of information they comprise, digital libraries are often considered to have become one of the major web services (Liaw & Huang, 2003). They are also assumed to be among the most complex and advanced forms of information systems, and interoperability across digital libraries is recognised as a substantial research challenge (Gonçalves et al., 2004; Candela et al., 2007). Moreover, it is commonly expected that the today's library, archive and museum services will converge in the future digital content repositories (Marty, 2008). Most of the current research activities in this area relate to metadata object description, inter-object relations (semantic similarity, citation references, near-duplicates identification, classification), text and data mining and automated content processing, user personalisation and community services, and large-scale distributed architectures and infrastructure interoperability and performance. In this chapter, we will

analyse several examples of the current state-of-the-art in digital library and repository infrastructure technologies. Only the very recent years have seen the rapid increase of the pace of research and development of adequate technologies necessary to build large scale knowledge management and content provisioning infrastructures to support the individual advanced digital libraries and repositories and the associated automated content analysing systems. In order to be able to find a common ground for evaluation and comparison of different solutions, a universal formal description framework is required. While there is still no single formal digital library reference model in wide use, several approaches have been recently proposed, most notably Streams, Structures, Spaces, Scenarios, and Societies (5S), DELOS Digital Library Reference Model, and MPEG-21, although the latter, aimed at defining an open framework for multimedia applications, is not directly related to digital libraries. The Reference Model for an Open Archival Information System (OAIS) provides a framework to address digital preservation.

The 5S model, proposed in a PhD dissertation by Marcos André Gonçalves, introduces abstract concepts of streams, structures, spaces, scenarios, and societies providing means to define digital library objects, services and other entities. In the model, the streams are understood as simple sequences of arbitrary items used to represent serialized content, the structures are labelled directed graphs, organizing the streams, the spaces are seen as sets with associated operations on them, the scenarios are sequences of actions performed in order to accomplish functional requirements, and the societies are defined as sets of entities and activities, and the relationships among them. (Gonçalves et al., 2004)

Based on these abstract notions, the model proposes a formal ontology defining the fundamental concepts, relationships, and axiomatic rules governing the digital libraries. Contrary to other approaches, the 5S has the ambition to describe digital libraries in an axiomatic formal way. The model can be used equally as a base to build a digital library taxonomy, a quality model, or to perform a formal analysis of specific case studies. The basic concepts of the 5S models are summarised in Table 1.

Streams are sequences of elements of arbitrary types (basically bitstreams), representing serialised content objects (which of course may be text, being a stream of characters) or data transfers (like in streaming video). 5S differentiates between “static” streams, which simply correspond to stored data, and “dynamic” streams, which are data in transfer.

S5 defines s*tructures* as the means to organize and arrange components of an entity. The purpose of structuring a document is to orientate the reader in the information. A typical representation of a structure of a digital text object is its embedded markup (for example in an XML file). Similarly, relations or graphs structure raw data, or hyperlinks define a structure of a web site.

Scenarios are sequences of events denoting transitions between states of the system. They can be seen as ordinary use cases, describing desired external behaviour of the system from end users' perspective. They provide functional description of the system and therefore may be considered vital in the process of its design. Since the scenarios can be perceived as user-level service contracts, in many cases they may provide enough specification for system prototyping purposes. Each scenario describes a part of the system's functionality in terms of what happens to the streams, to the structures, and in the spaces through a sequence of events. Scenarios allow to quickly comprehend the complexity of a digital library and they are a common way of specifying system's functional requirements in its design phase. Moreover, the scenarios are one of the most intuitive ways of describing the system's behaviour.

Models	Primitives	Formalisms	Objectives
Stream Model	Text; video; audio; software program	Sequences; types	Describes properties of the DL content such as encoding and language for textual material or particular forms of multimedia data
Structural Model	Collection, catalogue; hypertext; document; metadata; organizational tools	Graphs; nodes; links; labels; hierarchies	Specifies organizational aspects of the DL content
Spatial Model	User interface; index; retrieval model	Sets; operations; vector space; measure space; probability space	Defines logical and presentational views of several DL components
Scenarios Model	Service; event; condition; action	Sequence diagrams; collaboration diagrams	Details the behaviour of DL services
Societies Model	Community; managers; actors; classes; relationships; attributes; operations	Object-oriented modelling constructs; design patterns	Defines managers; responsible for running DL services; actors, that use those services; and relationships among them

Table 1. The 5S Digital Library Model (source: Wikipedia)

A digital library's role is to serve the information needs (collecting, preserving, sharing, etc.) of its *societies*. Therefore, a society can be seen as the highest-level component of a digital library. In the 5S model, it is defined as a set of users, computers or software and the relationships among them and between them and their activities. Examples of specific human societies in digital libraries include learners, teachers, patrons, authors, publishers, editors, maintainers, developers, and the library staff. The traditional role of hardware and software members of digital library societies have been to support and manage services used by humans, but recently, they can increasingly be perceived as the users themselves (e.g. processing content served by another software). Also societal governance issues, including policies for information use, reuse, privacy, ownership, licenses, access management, and information integrity, are of fundamental concern in digital libraries.

A *space* is a set of objects complete with operations on them that obey specified constraints. It defines logical and presentation views of several components. The concept of spaces is particularly useful because of the generality of its definition. It can be used when a feature of a digital library cannot be represented by any of the other four basic concepts of the 5S model. A space, as defined in the 5S, corresponds to a mathematical notion of a space (including specific cases such as topological, metric, linear or vector space). A document space or a virtual collaboration space may serve as examples in the digital library domain.

The DELOS Digital Library Reference Model (Candela et al., 2007) defines a three-tier digital library domain view, differentiating between a digital library, being the final system actually perceived by the end-users as being the digital library, a digital library system,

being the deployed and running software system implementing the digital libraries, and a digital library management system, being the generic software system supporting the production and administration of digital library systems and the integration of additional software offering more refined, specialised or advanced facilities. It also defines a number of digital library domain entities and relations between them, such as content, users, functionalities (actions), polices, etc. The DELOS Reference Model seems to be primarily focused on describing an autonomous digital library and does not cover the interoperability in a distributed environment.

The MPEG-21 standard, from the Moving Picture Experts Group is ratified in the standards ISO/IEC 21000 - Multimedia framework (MPEG-21). Its primary purpose is to define an open technology framework needed to allow users to exchange, access, consume, trade or manipulate multimedia in an efficient and transparent way. The standard is based on a definition of a Digital Item (DI), as a fundamental unit of distribution and transaction. A Digital Item is defined as a structured digital object with a standard representation, identification and metadata. The Digital Item is the digital representation of an asset and an entity that is acted upon within the MPEG-21 framework. Parties that interact with the Digital Items in the MPEG-21 environment are categorized as Users acting in different roles. The aim of MPEG-21 is to provide a set of tools allowing Users to interact between themselves and the objects of that interaction are Digital Items. These User-to-User interactions may include providing, modifying, archiving, or consuming content etc. In order to allow various parts of the standard to be used autonomously, MPEG-21 is organized into a number of independent parts, including: Digital Item Declaration (DID), Digital Item Identification (DII), Intellectual Property Management and Protection (IPMP) for license enforcing, Rights Expression Language (REL), Rights Data Dictionary (RDD), Digital Item Adaptation (DIA), or Digital Item Processing (DIP). While originally promoted for the media industry, mostly for its strong support of Digital Rights Management (DRM), it has also found some ground in the digital library domain, namely the aDORe project developed by Los Alamos National Laboratory (Van de Sompel et al., 2005).

2. Historical examples

The concept of a distributed content provisioning infrastructure is nothing new or unique to digital libraries. The Internet has seen successful, truly distributed, not centrally managed, large scale content infrastructures in operation for years. Internet Usenet, developed from the general purpose UUCP architecture (Novitz & Lesk, 1978) in the early 1980s, and still in massive use, may be one example here. Usenet is distributed among a large, changing and evolving conglomeration of servers that are loosely connected in a variable, yet robust mesh. Its servers store and forward messages to each other, and also provide read-write access to the clients (Salz, 1992). The servers may act in various roles, such as feeders, stampers, transit or storage servers. At the same time the Usenet content is structured into a hierarchy of groups with delegated management authorities. Current Usenet contents typically include text messages, images, computer software, as well as other multimedia objects. Infrastructure-wise, Usenet is governed by a set of protocols for generating, storing and retrieving its contents and for exchanging it among its widely distributed readership (Horton & Adams, 1987). The backend Usenet infrastructure employs the peer-to-peer architecture, that was rediscovered under that name only years later.

Another example of a truly distributed, massive information management system may be Archie, considered to be the first Internet search engine (Sonnenreich, 1998). Created in 1990, before the World Wide Web, Archie was the Internet search engine for the FTP sites' contents. While not directly dealing with content delivery, Archie focused on harvesting and indexing the content and providing search service. Similarly to Usenet, the independently operated, heterogeneous Archie servers, together with the archives' mirrors system created a complex grid-like infrastructure, in which they exchanged data about the harvested FTP servers and provided services to the users, several years before the "grid" term was coined and grid computing reinvented.

A more sophisticated example may be the Harvest Information Discovery and Access System, developed in 1994 (Bowman et al. 1995) part of which evolved over time into Squid Proxy Cache (Wessels, 2004). The Harvest infrastructure consisted of several, possibly replicated, distributed components: gatherers, brokers, indexers, replicators and caches. The gatherers, in their basic role being simply web robots, shared also some characteristics with the modern OAI-PMH servers/harvesters. At the time of the Harvest's implementation, there was still no HTTP/1.1 and the If-Modified-Since GET was not widely implemented. The Harvest system proposed that HTTP server managers (the providers) run gatherers periodically against their public contents, and this way maintain incrementally updated content summaries, bypassing the need of futile queries downloading all the contents, regardless if they changed or not. The brokers would then retrieve the information from the gatherers (or other brokers) through SOIF protocol, and would invoke the indexers through a unified interface to index them. The interface would allow for different broker and indexer implementations to communicate. Harvest provided also a weakly consistent, replicated wide-area file system called mirror-d, on top of which brokers were replicated. Finally, Harvest included a hierarchical object cache, with each cache server communicating with its neigbours and parents with an ICP protocol (and later through cache-digests). The Harvest system provided a complete, scalable, distributed content delivery and replication infrastructure.

The early digital library systems were largely centralized monolithic databases or search engines with minimal scalability, focused on providing access to bibliographic information. Often originating from libraries' online public access catalogues (OPAC), they soon were reaching their expandability limits. The early digital libraries were usually closed systems, accessible only to human users, through either text terminals, Z39.50 protocol or through their specific web user interfaces. In their evolution they changed from single library systems to large federated digital libraries, but their basic constraints remained.

While the resulting single-purpose centralized institutional systems were adequate for their intended usage, they also proved useful for examining issues such as better understanding what functionality a digital library should possess, and determining which interfaces users find most appropriate. At that stage, performance and scalability were still of secondary concerns.

An alternative approach was assumed by the digital libraries originating from open access repository systems (institutional, thematic etc.). The repository systems focused on storing and making available for download full texts, usually of research papers, either published or pre-prints, most frequently represented in PDF. A typical early open repository did not offer, or offered only limited full text search capabilities but its metadata were made available for batch downloading through an OAI-PMH interface, thus enabling federated

searches to be performed by meta search engines *(Pieper & Wolf, 2007*), and providing a use case for future distributed infrastructures development.

3. Digital library infrastructures

As already stated in the introduction, it is believed that digital libraries are going to be among some of the most complex large-scale system infrastructures of the future. Modern digital library infrastructure systems feature service oriented multi-tier architectures with a loose coupling of modules. This component-based approach allows tailoring of individual deployments through the selection and replacement of required modules. Components are more natural units and easier to reuse than complete monolithic implementations. They also provide an alternative pathway to digital library federation and scalability, as distributed implementations are easier to implement with components running autonomously on different machines.

It is observed that, while the basic textual information search and retrieval techniques have been already mostly mastered, relatively well understood and implemented, many of the current challenges lie in generic data reuse and the associated methodologies. While it may seem simple at the first glance, the mere diversity of the possible data types and structures, not to mention the different access methods, make it an enormously complex problem. Also, knowledge representation, while in many aspects already addressed in theory (e.g. semantic network concepts or topic maps representation), is still in its infancy as far as the practical large scale usable implementations are concerned. Besides, knowledge discovery and extraction techniques, finding interrelations between heterogeneous objects of often different provenance, similarity analysis and compound objects handling are still not standardized. An additional challenge is posed, surprisingly, by the increase in the growth of the volume of scientific output. It is believed that in the not so far future, machines and automata will become the primary consumers of scholarly publications, as the quantity of produced information will sooner or later render humans incapable of effectively absorbing it without automated assistance. Therefore, already now it is anticipated that the knowledge, whether represented in the form of traditional publications, data, or more complex relations thereof, should be stored primarily in machine-friendly formats to best allow for its subsequent mass processing.

In general, the two primary challenges of all large distributed digital library infrastructures are the requirement to integrate the heterogeneous data and the system's true multidimensional scalability. Both are the necessary prerequisites allowing for subsequent efficient processing and analysis of the distributed content. Scalability is the base feature on which other desired qualities of a digital library system depend.

Scalability, as a general property of systems, is difficult to define (Hill, 1990). Traditionally, it is understood as the system's ability to be enlarged and to handle increased load in a graceful manner (Bondi, 2000). In the context of digital library systems, scalability in a multitude of dimensions is required, not only limited to the system's performance but also its extensibility and manageability. In order to fulfil the evolving requirements, and at the same time to remain competitive on a functional level, any large scale digital library system has to be based on a dynamic framework, undergoing constant development.

In a large scale digital library context, the system's extensibility is achieved primarily through the infrastructural approach. A distributed open infrastructure allows for a multidimensional scalability by a modular system's design, where different functionalities

can be realised through implementation of new or alternative modules. In a large scale distributed environment, the communication and overall management may become an issue. Distributed, component-based architectures are obviously more scalable than monolithic architectures. With a component-based approach, it is be possible to install a simple digital library system quickly and inexpensively on a commodity hardware. At the same time it is possible to deploy a complex system with custom functionality, high availability, and a replicated, distributed architecture within the same infrastructure. A digital library, requiring a specialized capability not supported by the system, needs to customize only the adequate components, and can reuse the bulk of the infrastructure without modifications. At the same time, a digital library without the need for a particular feature can omit components for that service in its deployment. A component corresponding to a particular performance challenge can be upgraded, replicated, or distributed, with minimal modification elsewhere in the system. A component-based approach also proves advantageous with heterogeneity issues, that can equally be present in content types but also in capabilities or search mechanisms.

While a Service Oriented Architecture (SOA) allows to build firm and extensible infrastructure systems, it imposes certain overhead both in the development cost and in the system's (communication) performance. This is alleviated by a more lightweight Resource Oriented Architecture (ROA) approach, which generally reduces the time to implement a system and also in many cases may result in a lower communication overhead.

Resource Oriented Architecture, however, does not offer true scalability, and may render large scale systems continuous development difficult to manage. To this end, the best solution may be a hybrid approach, offering well architected and tested stable SOA for the core services in the backend, and ROA for more rapid implementation of the front end web based services.

An early Open Digital Library framework has been proposed in 2001 (Suleman and Fox, 2001), in an attempt to define component interfaces for functions such as searching, browsing, combining metadata of different provenance, reformatting metadata, or providing a sample of recently added items. A prototype implementation has been prepared and successfully deployed.

aDORe is an infrastructure system developed at Los Alamos National Laboratory aimed at managing collections of objects stored in OAI-PMH enabled repositories, and making them available to external applications. The objects are represented in the system in the MPEG-21 Digital Item Declaration Language (DIDL) format. The Digital Objects in aDORe can consist of multiple datastreams as Open Archival Information System Archival Information Packages (OAIS AIPs), stored in a collection of repositories. The location of the repositories is kept in a Repository Index and the identifiers of each OAIS AIP, its represented object, and the relevant OAI-PMH repository where the object is stored, are contained in an Identifier Locator. The Identifier Locator is typically populated through OAI-PMH harvesting. An OpenURL Resolver provides OAIS Result Sets (presentable digital objects) to NISO OpenURL requests, and an OAI-PMH Federator exposes aDORe OAIS Dissemination Information Packages (OAIS DIPs) to OAI-PMH harvesters. Some concepts of the aDORe architecture may seem to resemble the Object Brokers of the Harvest system. Notably, aDORe makes an extensive use of MPEG-21 specification, which is rather unusual for a digital library system, as the standard seems to be mostly promoted by the media industry, interested in its DRM capabilities. The distributed storage in multiple OAI-PMH repositories

should make aDORe a relatively scalable system on the storage level. While the system is basically intended for local deployment, its modular architecture should also make it easier to be implemented in a distributed environment. While a centralised registry in the form of Identifier Locator may seem to create a bottleneck and a single point of failure, the system is capable of supporting tens of millions of documents. Nevertheless, generally, the component-based design of aDORe makes it possible to migrate between different implementations of the software modules without affecting the overall system's functionality. (Van de Sompel et al., 2005)

SeerSuite is a set of tools constituting a framework of an academic digital library built automatically by retrieving scientific contents found on the Web. SeerSuite tools are used by a couple of Internet services, most notably by CiteSeer^{X}, an index of publications in computer and information science and related areas such as mathematics or statistics, comprising over one million objects (Teregowda et al., 2010). The tools support full text indexing of the harvested contents and automatic citation extraction, indexing and linking. The basic components of the suite include a crawler, text and metadata ingestion and extraction tools, XML and fulltext repositories, object and citation databases, full text index, user interface, personalisation database and workflow supporting scripts. One of the design goals of SeerSuite, replacing a previous CiteSeer software, was a possibly high level of content processing automation. Once found by the crawler, a research paper, usually in PDF or PS format, is harvested and its text payload is extracted and analysed. At this stage, the text is being filtered to avoid indexing non-academic documents, and metadata, including citations are automatically recognised and extracted. The document is assigned a unique identifier, and duplicates are identified and handled. All the generated information is stored either in a database or in a form of XML in the repository. Also a copy of the original retrieved document is kept, and the citation database is updated accordingly. The files in SeerSuite are versioned and time-stamped and the full text index is incrementally updated taking this information in the account. This approach allows to avoid costly rebuild of the whole index each time a document is added or changed. Independently, MyCiteSeer portal keeps user profiles, portfolios and queries and supports building private collections, social bookmarking, user alerts and other similar personalised services. Individual services exchange data access object (DAO) information, or communicate through SOAP or REST interfaces.

A differentiating factors of SeerSuite include extensive metadata extraction tools and a strongly synchronised standalone citation graph service. While the system is very focused and remains centralised, a notable design effort has been taken to decompose it into a collection of autonomous tools (services) that can be potentially used on their own as building blocks of a future distributed digital library infrastructure.

A more universal infrastructure system, YADDA (Yet Another Distributed Digital Archive), designed along the lines of open knowledge environment paradigm, originated as a replacement software for Elsevier's ScienceServer platform. To this end, not only the extensibility but also high performance and high scalability were among its main design goals. For a number of years, ScienceServer was the primary (and the only) platform providing online access to journals to Elsevier's subscribers. In this time its one instance provided access to several million fulltext articles from Springer and Elsevier to all Polish academic and research institutions. Elsevier's announcement to terminate the development and support of the platform led to the necessity of looking for an alternative solution. It was

decided that a new, open system would be developed, not only meeting the functional and performance requirements of the high traffic journal provisioning platform but also capable of supporting in house developed bibliographic databases and repositories, open access content, books and other media, and integrating them in a single unified point of access for the end users. (Zamłyńska et al., 2008)

Contrary to many other digital library management systems, the YADDA suite models a much broader environment beyond the simple content items, and YADDA objects equally include object hierarchies, compound objects, actors, roles, licenses or institutions, and relations between them.

The basic YADDA environment employs the web services framework acting as a collection of APIs to services that can be accessed remotely. YADDA infrastructure consists of a set of core services including Object Storage Service, Metadata Storage Service, Structured Browse Service, Index Service, Workflow Manager Service, and AA Service and a number of extension services.

The Object Storage Service intended to store large volumes of mostly binary data supports full synchronization and versioning. In addition to that, it supports hierarchical data storage, in a manner similar to a traditional filesystem. Specific backends of the Object Storage Service allow to access objects using either YADDA-specific optimised interfaces, or well-known standard protocols like FTP, HTTP, or rsync.

The Structured Browse Service is an OLAP cube concept based module for managing relations between stored objects. The service allows to define relations and to query their data. It furthermore supports a number of specific non-standard field types such as enumeration string fields or bit sets with fast mask queries, which are particularly useful in the case of license credentials. The service allows for effective querying of aggregated data, or fetching the count of objects fulfilling specific search criteria. It supports lazy materialization of aggregated views, in which the results of predefined queries are materialized and the materialized tables are updated when the contents are accessed. The service also allows to define indexes on both relations and aggregated views.

The Index Service provides a flexible, fast and effective full-text search capability without restrictions on the type of the indexed documents. Depending on a particular setup, a number of Index Service instances can co-exist simultaneously, for scalability, load balancing, or reliability purposes. Index groups can be defined and searched in a single query. The service is transactioned, and its performance can be improved by splitting the index and/or storing it in memory. The service provides effective iteration through search results and filtering of frequent logical conditions in queries. Frequently executed, big boolean queries can be defined as filters which, when used, speed up searching up to 10 times. Currently, two different implementations of the YADDA Index Service API exist, with different functionalities, based on Lucene and SOLR, that can be used interchangeably.

The Workflow Manager Service is a subsystem responsible for scheduling and executing predefined tasks on the objects stored in the repositories. The tasks are organized in "processes", which define the sequences of the events. A process consists of nodes, each being a relatively simple operation awaiting an input and producing its output. During its execution, a node can access other YADDA services, and invoke associated actions. A simple example of a process node accessing a service may be a metadata reader, which takes an object's ID as an input, queries the Metadata Storage Service, and provides the object's content as the output. Sets of predefined nodes can be configured into chains and executed.

Processes can be run manually, can be scheduled or can be triggered by operations on other services, particularly by changes in the Metadata Storage Service.
The Metadata Storage Service (formerly the Catalogue service) is primarily responsible for storing rich metadata. This service provides synchronization, version control and search for metadata objects meeting specified criteria. A number of processes, defined in the Workflow Manager make use of the Metadata Storage Service data, including:

- A general indexing process, retrieving object hierarchy information (for example an article belonging to a volume of a journal published by a publisher) from the metadata structure elements and storing it in particular relations of the Browse Service (hierarchical relations and contributor-publication relation) and in the fulltext index.
- A metadata extraction framework, which runs as a multi-level process. First, a PDF file or an image is converted to a set of characters with assigned locations through optical character recognition. Next, the page layout is discovered and finally particular zones are tagged as title, author, abstract, keywords, references, etc.
- Citation parsing and matching by a rule-based citation parser. A network of citations is created by matching parsed citations with entries in the repository.

The Authentication and Authorization Service is designed as an open and distributed system, providing sophisticated security that allows to support a network of repositories and clients. It implements a complex yet transparent authentication and authorization layer based on XACML and SAML standards. One of the service's most significant features is the separation of authentication, authorization and policy enforcing functions. Thus it is possible to separate authority providers (users databases, client institutions etc.) and content providers (repositories which rely on the authentication data provided by authority providers, and which serve particular content). Furthermore, the service allows to propagate trust relationships in the network of repositories and clients (so-called "webs of trust"). Since the service uses XACML as a policy definition language, it is possible to define a variety of rule-based access policies in a flexible way. Each YADDA service supports Authentication and Authorisation Service based security layers, which allows to assign specific licenses to each object maintained by these services. Using XACML, it is possible to define flexible ways of limiting access to all objects according to their particular licenses.
Besides the core services, the YADDA environment contains a number of optional extension services, including a categorisation service, a similarity service, a citation extraction service, a reference service (citation graph and index) and a choice of interface services, including web GUIs. A standalone tool, DeskLight (being in fact a YADDA instance itself) allows for content publishing and online or offline collaborative content curation. All YADDA services and tools, particularly the YaddaWEB user interface and the DeskLight application are fully multilingual – with full Unicode and left-to-right and right-to-left writing support. The underlying data model allows to maintain multilingual information about any given element. For example a single publication can have its corresponding abstracts or keywords in a number of languages at the same time.
A number of tools have been developed for loading and bulk converting imported data from proprietary formats to the internal YADDA format or to export the data using standard formats and protocols (like OAI-PMH).
Formal service contract definitions allow user-specific security to be introduced to any service. Repository descriptors in the form of XML files provide descriptions of all services available in a given repository, allowing automated discovery and connection. Besides, the

service contract definitions allow to automate the service concertation process, service conformance testing and troubleshooting.
A proof of the YADDA environment flexibility and its down-scaling capabilities may be its embedded instance, DeskLight, which consists of custom lightweight implementations of the core APIs together with a couple of specialized editing tools and a GUI, all packaged as a java application intended for desktop use. DeskLight may be used as a local metadata editor, synchronizing the data with other DeskLight and YADDA server instances, thus allowing for efficient collaborative editing.
YADDA is a remote facade based service system, rendering it indifferent to the underlying inter-service communication protocols. The approach allows the services to be easily used in different deployment scenarios, ranging from tightly-coupled high performance scale-up installations to extensive, large open standards based distributed systems with service-level redundancy. The resulting flexibility of YADDA allows for its various components (services) to be easily included individually or in groups in other digital library infrastructures. The feasibility of this approach has been confirmed by diverse employment of various YADDA components in a number of different systems and environments.
Besides the original Elsevier and Springer journals application, individual YADDA services have been used in a number of different deployments, four of which are briefly presented below: DRIVER's Network Evolution Toolkit (used in several individual installations itself), OpenAIRE service, European Digital Mathematics Library, and BazTech database.
D-NET (DRIVER Network Evolution Toolkit) is a Service Oriented Architecture (SOA) based software suite created for the DRIVER digital library, aggregating the contents of the European research open repositories. The web services based suite allows to build a distributed infrastructure composed of a number of services, including an index, browse, store, OAI-PMH, collection, transformation, similarity, citation, text engine service and a number of D-NET specific orchestration services such as authorisation and authentication, information or manager service. Notably, version 2 of D-NET supports compound objects handling. Depending on a particular instantiation of the software suite, D-NET services can be combined into larger applications. The same services can be also shared among different environments. Individual services active in a D-NET instance register with its Information Service, allowing other services to discover them. The D-NET system's workflow is managed by a dedicated manager service responsible for executing other services in a desired sequence. (Manghi et al., 2010; D-NET: release of the DRIVER Software, http://www.driver-repository.eu/D-NET_release) D-NET successfully employs a number of YADDA infrastructure components, including its index, object store, authorisation and authentication, citation and referencing, and similarity services.
Another digital library system, where YADDA modules are being used is EuDML - the European Digital Mathematics Library (Sylwestrzak et al., 2010), currently in prototype. The EuDML system will consolidate the European information space in mathematics, harvesting national and local digital libraries and repositories and unifying and enhancing their metadata. The system, which will follow Service Oriented Architecture, will reuse existing technology but also develop new modules acting as services. The EuDML background services will include metadata harvester, registry and conversion manager, storage, search and browse, AA, and workflow manager. Besides the core, there will be a number of enhancement tools and services including citations manager, content annotation, author matching, data enrichment, personalisation and user interface with accessibility features. EuDML will use structured browse, index, storage, AA and citation services from the

YADDA environment. It will also use REPOX and MDR services developed for Europeana for metadata harvesting, mapping and managing. (Reis et al., 2009) The primary design goals of the EuDML platform are its extensibility, allowing easy addition of new services (and content), and its scalability in many dimensions, including the content's volume, content's structure, number of services, number of concurrent users, etc., without performance or reliability degradation. To this end, the system will be designed in a modular, distributed architecture, allowing to replace, upgrade or provide alternative services realizing the same or similar functions in the future versions.

OpenAIRE is a European initiative to provide an open-access publication repository infrastructure for scientists conducting research fully or partially funded by the European Commission. It is intended that, after leaving its pilot phase, OpenAIRE will provide an infrastructure to mandate open-access to all output of any research funded by the European Union, including textual publications but also data and multimodal results. Similarly to DRIVER, OpenAIRE uses selected YADDA services, including the Object Storage, Index and the Authentication and Authorisation Services. Users can upload their publications either to a central OpenAIRE repository (run by CERN), to the supported thematic repositories, or to their local open-access repositories and register the upload with the OpenAIRE system through a portal available at http://www.openaire.eu/.

A different application scenario for YADDA is BazTech - the citation database of Polish research journals in technology and related disciplines. While the BazTech database is centralized, its creation and updating process is highly distributed, and organized in a hierarchical manner. BazTech is maintained by a consortium of the libraries of Polish technical universities. In each library, its employees update the data in a local copy of the repository. The metadata are edited and the fulltexts uploaded using the DeskLight version of YADDA. The new contents are supervised, and when approved, the local repositories are merged together to form the eventual central BazTech database, running on another YADDA instance. Similar YADDA setups are used by a number of other project with similar usage characteristics.

The diversity and the multitude of different YADDA services deployment scenarios may serve as a proof, confirming that an open digital library service infrastructure concept is feasible not only as a prototype but also it excels in real life heavily used production systems.

4. Conclusion

Digital libraries related technology has undergone significant changes in the recent years. While the evolution path from the simple, autonomous, single-purpose monolithic systems towards multi-tier open infrastructural solutions may seem obvious, a lot remains still open for future research and subsequent development. There is yet no single widely adopted and mature enough production quality solution that would fully warrant adequate development potential beyond the immediate needs. In fact most of the currently deployed solutions constantly lag behind the requirements and expectations. Similarly, there are no well established flexible, performant and scalable digital library service to service communication standards, besides the basic protocols mostly pertaining to metadata transfers.

Besides the technology, also our understanding of user-centric design approach changes from the initial perception that service consumers are human actors towards seeing them

increasingly as other services processing the available textual or digital data and generating new semantic knowledge and pieces of information. The key to a successful and future-proof digital library system seems to lie in basing it on a standardized, open infrastructure that would be able to adequately expose content for automated machine-processing, much of which remains yet to be seen.

5. References

Alexander Ivanyukovich, Maurizio Marchese, Fausto Giunchiglia, (2008). ScienceTreks: an autonomous digital library system. *Online Information Review*, Vol. 32 Iss: 4, pp. 488-499

Bondi, A.B. (2000). Characteristics of scalability and their impact on performance. *Proceedings of the 2nd international workshop on Software and performance*. Ottawa, Ontario, Canada, 2000, ISBN 1-58113-195-X, pp. 195–203

Bowman, C.M.; Danzig, P.B.; Hardy, D.R.; Manber U., & Michael F. Schwartz M.F. (1995). The Harvest Information Discovery and Access System. Computer Networks and ISDN Systems, Vol 28, Issues 102, pp. 119-125. doi:10.1016/0169-7552(95)00098-5

Candela, L.; Castelli, D.; Ferro, N.; Ioannidis, Y.; Koutrika, G.; Meghini, C.; Pagano, P.; Ross, S.; Soergel, D.; Agosti, M.; Dobreva, M.; Katifori, V. & Schuldt, H. (2007). The DELOS Digital Library Reference Model – 0.98, p. 20

Emtage A. & Deutsch P. (1992). Archie - an electronic directory service for the Internet. *Proceedings of the USENIX Winter Conference*, pp. 93-110, January 1992.

Gonçalves, M.A.; Fox, E.A.; Watson, L.T. & Kipp, N.A. (2004). Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Model for Digital Library Framework and Its Applications. ACM Transactions on Information Systems, 22, 2, (April 2004), 270-312, ISSN:1046-8188

Hill, M.D. (1990). What is scalability ? *ACM SIGARCH Computer Architecture News*, Volume 18 Issue 4, pages 18-21, ISSN 0163-5964)

Horton, M. & Adams, R. (1987). Standard for Interchange of USENET Messages, RFC 1038 (December 1987)

Liaw, S. S. & Huang, H. M. (2003). An investigation of users attitudes toward search engines as an information retrieval tool. Computers in Human Behavior, 19, 751–765.

Manghi, P.; Mikulicic, M.; Candela, L.; Castelli, D.; Pagano, P. (2010). Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System. D-Lib Magazine, 16 (3/4), March/April 2010, doi:10.1045/march2010-manghi

Marty, P. F. (2008). An introduction to digital convergence: libraries, archives, and museums in the information age. Archival Science. Vol. 8, No. 4 (December 2008), pp. 247-250, ISSN 1389-0166, Springer

Nowitz D. A. & Lesk, M. E. (1978) A Dial-Up Network of UNIX Systems, In: UNIX Programmer's Manual, Seventh Ed., Bell Laboratories, Murray Hill, New Jersey

Pieper, D.; Wolf, S. (2007). BASE - Eine Suchmaschine für OAI-Quellen und wissenschaftliche Webseiten. Information, Wissenschaft & Praxis (IWP), 58(3), 179-182, ISSN 1434-4653

Reis, D.; Freire, N.; Manguinhas, H., Pedrosa, G. (2009). REPOX: a framework for metadata interchange. Lecture Notes In Computer Science. Proceedings of the 13th European conference on Research and advanced technology for digital libraries. pp. 479-480

Salz, R. (1992). InterNetNews: Usenet transport for Internet sites. Proceedings of Summer '92 USENIX, pp. 93-98. June 8-12, 1992 – San Antonio, TX

Schwartz, M. F.; Emtage, A.; Kahle, B & Neuman, B. C. (1992). A Comparison of Internet Resource Discovery Approaches, Computing Systems, pp. 461-493, 5(4), August 1992

Sonnenreich, W. (1998). A history of Search Engines, In: Web Developer Guide to Search Engines, Sonnenreich, W.; Macinta, T., p. 464, Wiley, ISBN 978-0-471-24638-1

Suleman, H.; Fox, E.A. (2001). A framework for building open digital libraries. D-Lib Magazine 7(12), ISSN 1082-9873, available online at http://www.dlib.org/dlib/december01/suleman/12suleman.html

Sylwestrzak, W.; Borbinha, J.; Bouche, T.; Nowiński, A.; Sojka, P. (2010). EuDML—Towards the European Digital Mathematics Library. Proceedings of DML 2010. pp. 11-26, Paris, France (Jul 2010). ISBN: 978-80-210-5242-0

Teregowda, P.B.; Councill, I.G.; Fernández R., J.P.; Kasbha, M.; Zheng, S. and Giles, C.L. (2010). SeerSuite: Developing a Scalable and Reliable Application Framework for Building Digital Libraries by Crawling the Web. 2010 USENIX Conference on Web Application Development, June 23–24, 2010, Boston, MA, USA

Van de Sompel, H.; Jeroen Bekaert, J.; Liu, X.; Balakireva, L.; Schwander, T. (2005). aDORe: A Modular, Standards-Based Digital Object Repository. The Computer Journal. 48(5) pp. 514-535, doi:10.1093/comjnl/bxh114

Wessels, D. (2004). Squid. The definitive guide. O'Reilly and Associates ISBN 0-596-00162-2

Zamłyńska, K.; Bolikowski, Ł.; Rosiek, T. (2008). Migration of the Mathematical Collection of Polish Virtual Library of Science to the YADDA Platform. In: Sojka, Petr (ed.): Towards Digital Mathematics Library. Birmingham, United Kingdom, July 27th, 2008. Masaryk University, Brno, 2008. pp. 127-130

2

Law of Success or Failure in the High Tech Driven Market – "Revenge of Success" in the Biotech, Nanotech, and ICT Industry

Takayama, Makoto
Niigata University, Graduate School of MOT; UCLA Medical School
Japan

1. Introduction

In the case of product change, it is well known that incremental product innovation is well managed by the cooperation between marketing knowledge and technology knowledge (von Hippel, 1988, 2005, 2009; Clerk and Fujimoto, 1991). If knowing marketing needs and technology seeds is enough to develop the new product, the market leader with research capability could hold the best position to become a successor in the field. The reality is different from this assumption; the successor is very often replaced (Christensen, 1997).

There is a lot of discussion on success factors in product innovation: which is the determinant for success in product innovation, technology-push or market-pull? According to technology-driven theory, the importance of technological innovation is highlighted for product innovation (Rosenberg, 1976; Freeman, 1982; OECD, 1984; Dosi. 1982, 1984, 2010a, 2010b). In contrast, market-driven product innovation has highlighted the importance of market needs or customer needs (Rothwell et al., 1974; von Hippel 1979, 1980, 2005, 2009, de Jong & von Hippel, 2009). These discussions have not led to any sufficient conclusion. With respect to new product development (NPD), the aforementioned opposing theories have highlighted the importance of technology knowledge and market knowledge. Although the applicability of the theory depends on the respective product, it is absolutely obvious that the product is an embodiment of market knowledge and technology knowledge. In the case of novel technology emerging, technology seemed to serve as a major driving factor to introduce the relevant new product into a market. It is also well recognized that market knowledge can stimulate successful NPD. It seems to be obvious that market knowledge assists the accumulation of technology knowledge. Apparently, both technology knowledge and market knowledge are able to serve as the key factors that enhance successful NPD. If so, the market leader can keep its leading position in the market for the next generation of new products and hence the market leader cannot be easily taken over by the newcomer, even if the new product creates a new market. For continuous NPD, this assumption is true.

The concurrent engineering system is thought to be the best way to launch the next product into the market (Hammer & Champy, 1993). For the automobile industry and the electricity industry, many authors have demonstrated the advantage of the collaboration between

technology and the market (Ohno, 1988; von Hippel, 1988). The opposite of this finding is true. There are many examples of major players losing their strong position in the market when new products emerged. Why can't leading companies maintain the best position in the market? It seems to be related to the characteristics of the product. This problem has not yet been solved by the current discussions on NPD. In this chapter, a recent change to a product in the high tech markets such as bio, nanotech, ICT market are studied.
This work demonstrates that a strong product inhibits the NPD of a newly emerging product that will replace the existing product market in the future, furthermore, that there is no need to have the core competence for the existing market or technology. In order to demonstrate the behavior of the companies, R&D and market analysis have been carried out. Section 2 identifies the characteristics of product and players change in the high tech industry. Section 3 analyzes the behavior of the company in the case of product change. Section 4 describes the case studies of product change at mature stage of the existing markets. Section 5 presents mechanism of success or loss. Section 6 concluded the fate leads the success or loss of each of major players in the new born high tech market like invisible hand of God. How to evade from the fate is the most important implication at the end of this chapter; all businesses, though, cannot evade from the fate of the law of success or failure.

2. Current situations and structural problems of high tech industry management

The market of the high tech industry is in the process of development: the whole panorama remains invisible. Little attention has been paid to idiosyncrasies of the high tech industry in terms of technology and market. While it is relatively easy to forecast future development of existing industries, this is not the case for the high tech industry. The situation is analogous to the difference between physics and biology: logical forecasts useful in existing industries do not work for the high tech industry, which is as unpredictable as the world of bio-organisms.
Successful innovation and commercialization of new products or new business require a perpetual cycle of hypothesizing, verification and exploring about the products or businesses and their markets. Many enterprises have failed in high tech products or businesses because of this cycle conducted in a traditional manner. In other words, such enterprises were "revenged" by their success itself (Takayama, 2002, 2005; Takayama & Watanabe, 2002), only wasting their money, labor and time. There is a structure that prevents the conventional methods from succeeding, and the real problem is that not only major market players but also entrepreneurs are not aware of that structure set forth in the new born markets.

2.1 Absolute win or unavoidable loss in bio-industry businesses

As a matter of fact, many existing major businesses have failed in seemingly promising development projects especially of innovative products or businesses. Their extensive and preceding investments for R&D or facilities, including those for establishing new laboratories or huge infrastructures, did not prevent newcomers with different backgrounds from winning their market shares. The typical feature of the investment is prioritized and authorized by top management among the high tech related industries in common. Even with knowledge of new technologies, products and markets at higher level than their competitors, majors will certainly lose under some conditions, while winning under others.
As a typical case of the bio-industry market, in the applications of the recombinant DNA technique, all related firms had established bio-tech institutes under the prevailing bio-tech era

in 1980s. This bio-tech institute boom is not only limited to life science related firms such as pharmaceutical, beer, fermentation etc. but also absolutely unrelated firms in such industries as chemical, textile, food, steel, electrical industry etc. All businesses have believed still now that establishing only institute could strike gold mine from the huge unveiled markets.
The result of the win or loss was simple. Agricultural majors have continued to win the victory by successively launch of recombinant plant. They have kept competitors out of the existing market. Only the exception was a new born market segment like blue carnation and blue rose market made by outsider Suntory. On the contrary, pharmaceutical majors could not develop any bio-pharmaceuticals by own efforts. It should be emphasized that they could not take the opportunity by introducing bio-products even at very cheap licensing fee offered by bio-venture like Genentech, although business routine based on open discussion has been established in the organizational structure of every firm for a long time. Major players failed the bio-pharmaceutical market. Typical winner was Amgen that is started by spin-off researchers from Merck & Co. Inc. in the USA and became Board Chairman of PhRMA (Pharmaceutical Research and Manufacturers in America) and actual revenue including licensees is presumed almost equivalent to the sales amount of Merck. If Merck could hold scientists in its R&D activity, it could keep the steadfast and immovable position in the current market. Exception was catch-up type bio-products like insulin, human growth hormone. All of such catch-up type recombinant-products are marketed by strong players in the existing pharmaceutical market even if the first discovery was made by bio-venture like Genentech.
In conclusion, the fate of win or loss were opposite, as shown in Table 1: agricultural majors were successful in excluding newcomers from the recombinant plants market, while no pharmaceutical majors were able to commercialize recombinant drugs in spite of large-scale R&D programs in dedicated laboratories.

	Existing businesses' result	Winning business sector
Recombinant plants	Complete Victory	Agro-business
Recombinant pharmaceuticals	Unavoidable Loss	Pharmaceutical business

Table 1. Fate of major businesses in the new born market of recombinant bio-products

Meanwhile, the chemical industry, once seen as the leader in commercializing bio-products, has shown a general tendency of divestiture of bio-businesses, including the pharmaceutical business. Those new companies usually focus their resources in limited product areas. This results in cutthroat R&D competition in a small number of themes, naturally raising the share of R&D expenditure. This in turn necessitates pursuit of large sales by focusing on lucrative products, creating a vicious circle. Owing to such structural problems set forth in high tech innovation, all chemical firms have failed NPD not only in the bio-industry but also nanotech industry, which is demonstrated in the following sanction 2.2.

2.2 Absolute win or unavoidable loss in nanotech industry businesses

The fate of win or loss in nanotech products was the same. All firms in the nanotech-related industry have still now been establishing nanotech institute from 1990s. Nanotech boom started in 2000 and nanotech has prevailed over the material industry. Although nanotech broadly includes biotech, pure bio industrial area should be excluded from nanotech since bio started as recombinant technology earlier than nanotech in 1970s. Based on this definition, nanotech has been mainly applied to device and material.

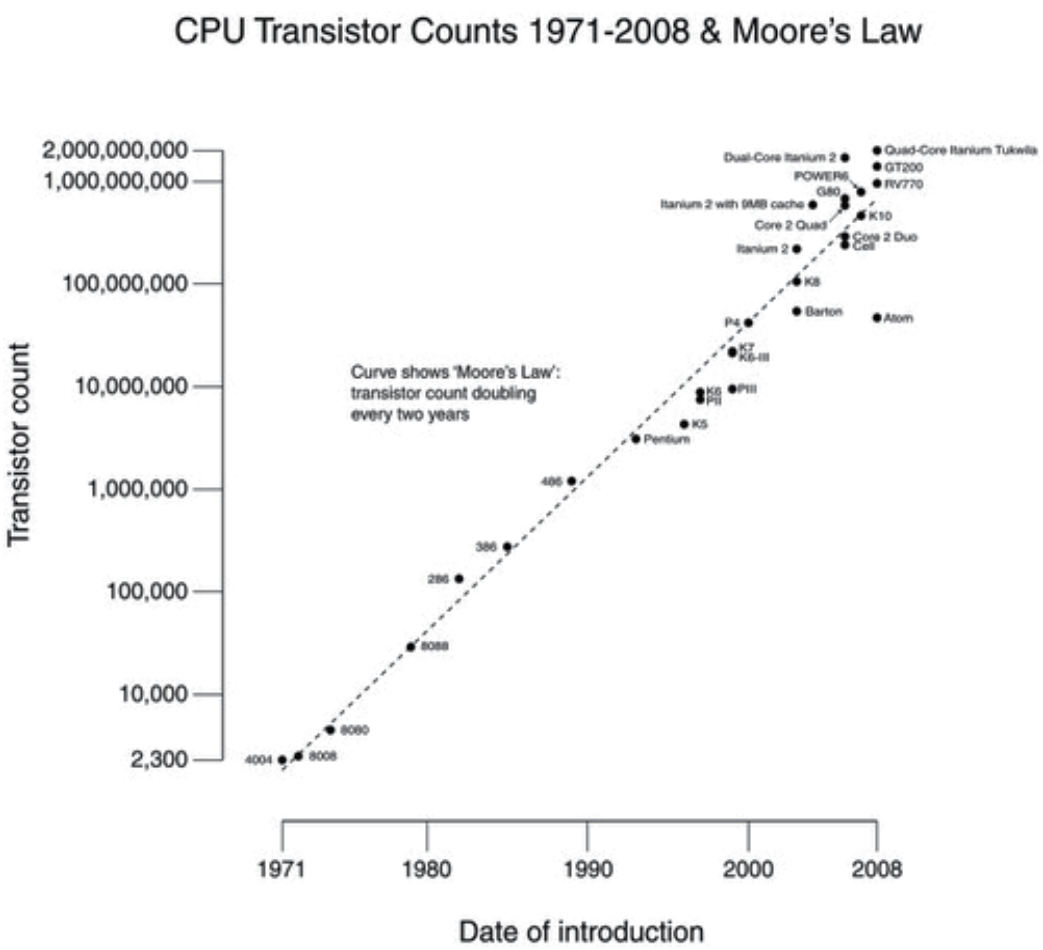


Fig. 1. Moore's law from Intel's history (Intel HP)

Exemplary examples in the nanotech device market are ICT devices like semiconductor, memory etc. The most famous law in this industry is "Moore's law" (Moore, 1965 and 1996). Moore's original statement was that transistor counts had doubled every year. As shown in Figure 1, Intel has succeeded in continuous innovation by improving the performance of integrated circuit. Sustainable growth of ICT has been suspicious since small scaling limit is believed to reach in 2012. Intel is overcoming this limit by applying nanotech to integrated circuit. This clearly proves that major player could keep the position in the next market by continuous innovation.

In case of nanotech material, the situation is completely opposite to the common prediction, as below. There is no noteworthy nanotech material market except carbon nano-materials like fullerene, carbon nanotube etc. A **fullerene** is any molecule composed entirely of carbon, in the form of a hollow sphere, ellipsoid, or tube as shown in Figure 2. Spherical fullerenes are also called **buckyballs**, and cylindrical ones are called carbon nanotubes or buckytubes. Buckyballs and buckytubes have been the subject of intense research, both for their unique chemistry and for their technological applications, especially in material science, electronics.

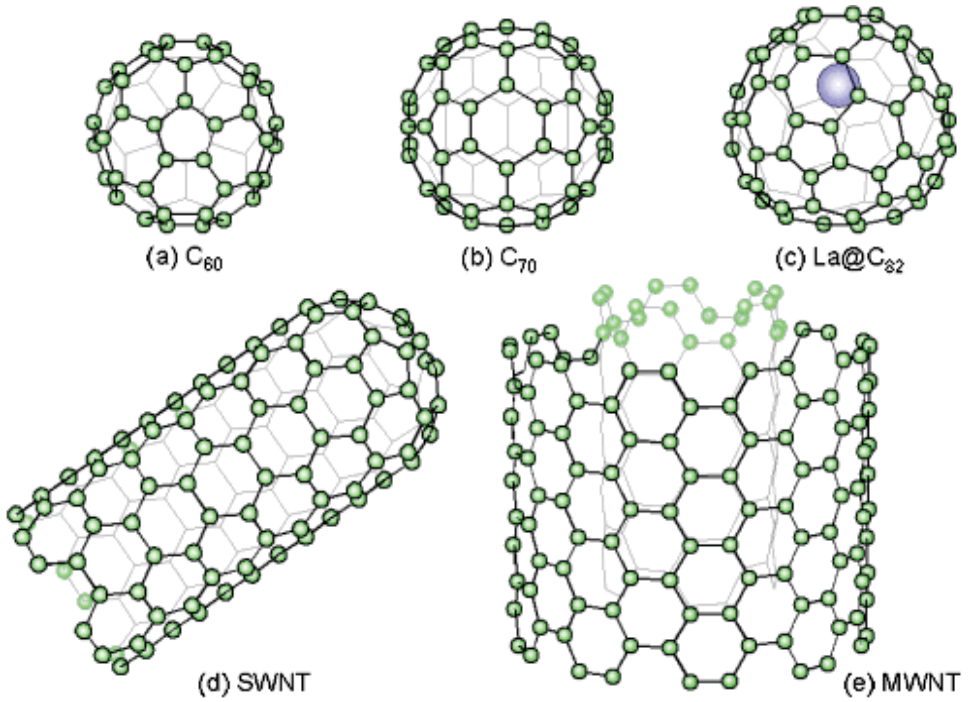


Fig. 2. Carbon nano-materials

The existence of C_{60} was predicted by Eiji Osawa in a Japanese magazine in 1970 (Ozawa, 1970). He noticed that the structure of a molecule was a subset of a soccer-ball shape, and he made the hypothesis that a full ball shape could also exist. His idea was merely reported in Japanese magazines. Also in 1970, R.W. Henson proposed the structure and made a model of C_{60} (Thrower, 1999). The evidence for this new form of carbon was very weak and was not accepted over 29-years period.

From the viewpoint of the required expertise in the chemical profile of nanocarbon, the chemical industry, once seen as the leader in commercializing nanotech products, has shown a general tendency of divestiture of nanotech businesses, although the cosmetics and textile businesses succeeded in marketing nano-materials. Major companies usually focus their resources in limited product areas. This results in cutthroat R&D competition in a small number of themes, naturally raising the share of R&D expenditure. This in turn necessitates pursuit of large sales by focusing on lucrative products, creating a vicious circle. The entry has not been still now made not by chemical firms. Owing to such structural problems, unavoidable loss happened like bio-industry. Mitsubishi Corporation, the largest Japanese general trading company did break into the first market entry in 2003. Mitsubishi Chemical, a chemical company in Mitsubishi group denied its market but was urged to follow a small portion of investment. In spite of such common recognition in the chemical firms, Mitsubishi Co. assured to succeed in nanocarbon materials independently from any chemical businesses (hearing from Kojima, appointed President & CEO and then currently Chairman, and project team at Mitubishi Co., 2002). This situation is absolutely the same as in bio-tech industry. Chemical businesses denied the market potential of the new market, although they were expected to work as key players for high tech innovation. Only outsider could see the market potential.

As summarized in Table 2, the fate of win or loss were opposite between device such as semiconductor and material like nano-carbon. IT device majors like Intel were successful in excluding newcomers from the semiconductor market, while no material businesses were able to commercialize nanocarbon materials except cosmetics and textile in spite of large-scale R&D programs in dedicated laboratories.

	Existing businesses' result	Winning business sector
Semiconductor	Win	IT business
Carbon-nano material	Unavoidable Loss	General trading business

Table 2. Fate of major businesses in the new born market of nanotech products

2.3 Absolute win or unavoidable loss in ICT industry businesses

Information and Communication Technology or ICT allow users to participate in a rapidly changing world in which work and other activities are increasingly transformed. ICT can be employed to give users quick access to ideas and experiences from a wide range of people, communities, cultures and political issues. In the recent decades widespread incorporation of ICT into many tiers of business and structuring of the global economy has occurred. ICT has increased international interconnectedness and sped up the process of globalization. In conjunction with globalization and the information revolution, ICT has reshaped the workforce and business system. By increasing the speed of international communication, ICT has enabled businesses to outsource jobs, both in the manufacturing as well as white collar sectors (Rice, 2005).

In accordance with diversification of communication method, this trend, ICT accelerates the business model from mortal to click. By structuring network among everything, ICT increases the accessibility to the necessary information and decrease of the transaction cost. This feature brought forth new economical role as intermediate. Critical changes by ICT owed to the increase of utility of information and therefore caused in changes of transaction systems.

Opposite to the semiconductor business, existing businesses such as securities, retailing, advertisement, music did not take initiative for the transaction business system by using ICT, as shown in Table 3. Beyond the common expectation, all the new business systems are taken over not by outsider but also by new comers. This is the typical feature of ICT business, which is not observed in other high tech field.

	Existing businesses' result	Winning business sector
Securities	Loss	Internet security business
Retailing	Loss	Internet shopping
Advertisement	Loss	Net businesses
Music	Loss	Download music

Table 3. Fate of major businesses in the new born market by using ICT

The case in the network music business was the same. The Network Music business is a rapidly emerging new market. Furthermore, the business approach has been changed frequently in response to rapid changes in customer preferences and the constant evolution

of technical platforms. Network Music systems are rapidly changing from one technology to another. In order to achieve optimal corporate profit, the most crucial factor should be "how to create strong customer relationships" through the network as a continuing source of network music business. Key success factors in the network music business are now understood to be the capability to continuously create new products based on the infrastructure of customer satisfaction in the network system. Before download music market, Walkman has prevailed across the world. Walkman was the strange tool because the use was limited only through earphone not by speaker. Download music player like iPod is much easier to store and select the music. Sony denied the market potential of download music at top management conference in 1999 (Idei, 2009). Furthermore, it is noteworthy that Japanese firms are most good at the technologies used in iPod. This proved that technological top tier is not sufficient for win the new market, in other words, not useful for successful launch of innovative product.

2.4 Structural problems of high tech industry management

These situations raise a question whether the typically research-intensive high tech industries like bio, nanotech and ICT, with by far the largest or larger R&D expenditure in the whole industrial sector, have succeeded in innovation through high tech. In case of pharmaceutical firms, while publicizing themselves as the leaders in bioengineering for healthcare, existing companies in this field are losing the new market to newcomers, and trying to counter by expansion through M&A, resulting in yearly changing sales ranking. Smaller enterprises are forced to focus on a limited product lines. This may bring about higher efficiency and profit for a short period, but its long-term effectiveness is questionable. The present work discusses the management of high tech businesses to show that circumstances exist:

1. where major businesses succeed in product innovation allowing no newcomers to participate in the new market
2. where they fail because of the "revenge of past success";
3. factors exist that decide the win or loss.
4. by elucidating the mechanism of the win and loss, it is demonstrated that the product development in the high tech business requires a management strategy different from that for other industry branches.

3. Win or loss in the development of high tech products and business

3.1 Win or loss in the launch of high tech products

The performance of major businesses in the development of new biotechnology-related products is shown in Table 4. All the majors in the agricultural products have successfully commercialized bio-products such as recombinant crops and remain as market leaders. The food majors have also succeeded in assimilating biotechnology for renovation of production processes and development of new products. In contrast, the pharmaceutical and chemical majors were unsuccessful in antibody formulations and other biologics, except for products earlier developed such as insulin or growth hormones in spite of almost frenzied effort, including establishment of new laboratories dedicated to bioscience. Chemical companies have also lost the potential market of nanotech products even after the winners from other fields. Situations were similar in a related area: leading manufacturers of syringed did not develop needle-free syringes used in administration of biologics.

	Expected majors' result
Agriculture	Win
Food	Win
Bio-pharmaceuticals	Loss
Antibody pharmaceuticals	Loss
Nanotech materials	Loss
Needle-free syringes	Loss

Table 4. Performance of major businesses in new markets by high tech products

Win in biotech products development seems to depend on several factors. Table 5 sets win case to loss case for clarifying market position of majors to high tech products. Situations of agricultural products and biopharmaceuticals described earlier may be analyzed in terms of a few aspects.

	Agricultural majors	Pharmaceuticals majors
Business result	Win	Loss
Competition	Direct	Indirect/neutral
New products	Replace existing products	Create new markets
New product development	Promoted	Neglected

Table 5. Position of majors to high tech upcoming products

New agricultural products, such as recombinant crops, are in direct competition with existing products and will replace them as far as the advantage of the new products are maintained. This prompts the market leaders to keep their position by developing new products instead of insisting on their existing product lines. They can exploit wealth of relevant information for their competitive advantage, leaving little hope of market entry for potential newcomers. In fact, examples of successful entry by newcomers are limited to those in niche markets neglected by the majors.

Reverse is the case for bio-pharmaceuticals and antibody pharmaceuticals. Leading manufacturers of first-generation bio-products such as insulin, growth hormones started the development and launch in the market once the technology has been confirmed. As the facts described, they have immediately followed the emerging high tech itself and furthermore master the production and marketing of the bio-pharmaceuticals. In spite of core capabilities, all of majors failed in development of granulocytic proliferation factor, or the multibillion-dollar erythropoietin, antibody pharmaceuticals, which compete only indirectly with existing products or are neutral to competition, and create their own new markets as shown in Table 6. The majors neglected products development because they failed to recognize the potentially huge market size for those products (which they predicted, instead, would form only small niche markets). In other words, the majors were not willing to be competitors in the new field, thus allowing newcomers to dominate the market easily.

This miscalculation was also responsible for their failure to respond to the need for marketing partners of the newcomers without established sales network, which would have meant an opportunity to seize on the new market without compromising the existing products. This is a typical case of the "revenge of success" (Takayama, 2002; Takayama & Watanabe 2002). Amgen, a pioneer of bio-pharmaceuticals, benefited from these

circumstances so much so that its top management provided a president of Pharmaceutical Research and Manufacturers of America.

	First-generation bio-pharmaceuticals	New born bio-pharmaceuticals
Business result	Win	Loss
Competition	Direct	Indirect/neutral
New products	Replace existing products	Create new markets
New product development	Promoted	Neglected

Table 6. Position of existing majors to first-generation bio-pharmaceuticals and new born bio-pharmaceuticals

3.2 Win or loss in starting new business system by using high tech

As a proof of the difficulties of the entry to business system by existing players, cloud computing exhibits many examples. Cloud computing had been recognized as next wave for technology investors (Hamilton, 2008). As Cloud platforms become ubiquitous, global cloud is expected to serve as an exchange and market infrastructure for trading services, since the need for internetworking create a market oriented global cloud exchange for trading services (Armbrust etc., 2009). SPI is KFS (Key Factors for Success) in cloud computing business; SaaS, PaaS and IaaS represent Software as a Service, Platform as a Service and Infrustructure as a Service., respectively. The sygnificant role of these core comcepts can easily make out the conlusion that the most critical key success factor for the cloud business is public cloud. In spite of such common recognition for the market size, only small firms like OpSource started to provide the public cloud service, since all majors have hesitated to use cloud computing system due to immaturity of the system itself.

4. Mode of competition as a decisive factor

Examples described above clearly shows that the mode of competition between new and old products is a decisive factor for corporate behavior with respect to new products. Direct or indirect competition of a new product with a company's existing products determines whether the company wins or loses in launch of that particular product or start of noble business system. Since the mode of competition of a new product depends on the nature of existing products, different companies may behave in different ways even in one and the same market segment.

4.1 Case of specified supplementary foods

Another example of such corporate behavior can be found in the rapidly growing market in Japan of "specified supplementary foods", a class of foods which contain specific therapeutic ingredients and are approved by the Ministry of Health, Labour and Welfare based on test results on the safety and effectiveness. A major area of such foods is life style related diseases as represented by metabolic diseases including diabetes mellitus, hyperlipemia, hypertension and hyperlithemia, which are the main targeted therapeutic area of all pharmaceutical businesses. Many supplementary foods against these diseases have been

developed by both big and small businesses, creating a market of about a trillion yen. It is predicted that "the market will more than double if the national health insurance system is modified to allow doctors to prescribe supplementary foods for prevention of adult diseases, particularly against fatigue and enhancing functions of blood vessel endothelium cells" (Nikkei Health, Jan. 28, 2005). For this purpose, the insurance system has begun to change from 2009. This is a considerable market size compared with 8,850 billion yen sales for all Japanese pharmaceuticals prescribed by physicians (mix, 2010) and 774 billion yen for OTC sold at drugstore (Yano Research Institute, 2009) in 2009.

Drug manufacturers are in the most advantageous position in this market sector with the expertise in new drug evaluation, familiarity with the Ministry's policies, and existing health food divisions that eliminate prior investments for distribution channels, as well as accumulated drug-related knowledge, experience and infrastructure that possibly bring about synergistic effects. In short, this market should be the easiest for drug manufacturers to attack. Actually, however, companies behaved as shown in Table 7. Drug manufacturers, including smaller ones, did not attempt product development, not to mention market entry. Food or drink majors behaved the same as drug majors. This is clearly not their failure but their intention. The new market was rapidly created in 2003 by catechin green tea drink of Kao, which was originally a soap company and currently a commodity major in the Japanese market. Only inexperienced firm has succeeded in entry to new born market. Similar phenomena have been observed in the American and European functional food markets. As recent topis, iPod by Apple was the same case, although the first winner in the former market failed to enter rather denied the market, as described in Section 2.3.

	Food or drink majors	Drug majors	Kao, outsider
Market entry	Loss	Loss	Win
Mode of competition	Indirect	Indirect	Inexperienced
New products	Create new market	Create new market	Create new market
Product development	Hampered	Hampered	Promoted

Table 7. Corporate behavior in supplementary food market

The catechin was well-known as bitter ingredient of green tea and lipid-lowering function for human body. As the most popular green tea drink major, Itoen Inc. discovered canned green tea in 1985 (Itoen HP) and had kept the top market share, around 40% for 20 years (Ishii, 2010). Drink major firms including Itouen, Suntory etc. know the mode of function of catechin for lowing lipid, they neglected as a transient boom even after Kao sold 20 billion yen after 10 months from the launch. More interestingly, all drug majors neglected the market itself notwithstanding exactly the same targeted as drugs. It is worthy of remark that winner has not held any value chain such as food or health-care channel in addition to lack of any authorized core competence for the new market.

The cases presented so far are concerned with new products that create new markets. But the same analysis can also be made for new products within existing market sectors. The same pattern of win and loss of the majors is observed in every country (Takayama, 2002; Takayama & Watanabe, 2002; Takayama, Watanabe & Griffy-Brown, 2002; Takayama, Takayama, 2005; Takayama 2009). Some cases are described in the following sections.

4.2 Case of the most competitive new product in the most competitive market

A typical example of such cases is the antihypertensive drugs, which accounts for about 10% of the world drug market from 1990s and now forecasted to increase the share in the market. Hypertension is a kind of lifestyle-related diseases, although few symptoms, heart failure, cerebral hemorrhage, myocardial infarction and other dangerous complications are caused.

The anti-hypertensive market is almost mature in the early 2000s, because the existing products treat almost 90% of patients. According to interviews conducted by the authors, only three companies in the top 20 pharmaceut ical companies in the world maintained research activity for hypertensive drugs in 1999 and the others have been winding down this activity, although all companies reinforced research activity on anti-hypertensive drugs at least 10 years ago. Last product innovation was emergent, although the anti-hypertensive market is in the mature stage. The final products, angiotensin receptor blocker (ARB; ATII: angiotensin II receptor antagonist) were made based on the same new technology and have been launched country by country. In Japan, the first product was launched in August 1998.

In hypertensive medication, there are two major products, Ca blocker (Ca) and Angiotensine Converting Enzyme Inhibitor (ACE). Since Ca shows rapid onset and sharp efficacy, it is used as the first choice for the treatment of hypertensive patients who do not have organ malfunction, such as diabetics. Although the efficacy of ACE is less than Ca and ACE has the side effect of a cough, ACE is used for older patients who are at risk from organ damage.

As the final new product in this market, the first product of ARB has launched by Merck Co. in 1995 and its peak sales was, at that time, estimated around 400 million dollars in the world. After 10 years of the first launch, this new product category has replaced Ca antagonist, the largest product category in the existing hypertensive market (Fuji Keizai, 2010). In 2004, sales of ARB in the Japanese pharmaceuticals market has exceeded hyperlipidemia market, which was the largest product category and also the main target of specified supplementary food as described in section 4.1.

After the severe competition of new drug development from 1990s to the early 2000s, have proved to dominate the world antihypertensive market with a share over 70% for the first prescribed patients. In the struggle for this huge market involving drug manufacturers of all sizes, the majors which had product lines not in direct rivalry with ARBs, e.g. calcium antagonists, were eventual losers: they did not succeed in development, if any, of the new product, at least in a timely manner. The "calcium myth", promoted in Japan by the leaders of the antihypertensives market, which claimed the superiority of Ca antagonists, did not play any important role in the process. Rather, the majors fell victim to the revenge of their own success in every country (Takayama, 2002; Takayama & Watanabe, 2002). The performance of specific enterprises is summarized in Table 8 by the top 10 companies in the world Ca market. Of the top 10 companies, nine have no ARB product and two firms with minor share did get the co-marketing right of ARB from other marketing partners. Although two companies, Takeda and Novartis, have ARB pro ducts, Takeda does not market Ca outside of Japan, and Sandoz Co. and Ciba-Geigy Co. (who are merged and became Novartis) brought ARB in 1997. Two companies, Hoechst and Astra, are developing license-in ATII products. Ten out of the top 10 companies have no self-made or self-developed products in the world market, although those products became the global mega breakthrough products in the middle of 2000s.

In contrast, most of the major producers of ACE inhibitors which are indirect competition with ARBs succeeded in product development (see Table 9) (Takayama, 2002; Takayama & Watanabe 2002). They were able to exploit their superiority in information access for more speedy development, which prevented effectively newcomers from entering the market.

The behavior of ACE leaders is different as demonstrated in Table 9. The positive behavior of leading companies4 for developing ARB or ACE/NEP. ACE/NEP is expected to be a superior product to ACE, like ARB, because it has higher potency than ACE and reduces the cough side effect of ACE by adding NEP inhibitor activity. Seven out of the top 10 companies are developing their own products and one company is developing a license-in product. This fact demonstrates the positive attitude of the ACE leader1 for developing ARB. The remaining two companies do not develop ARB. This is because of their strong position as first and second in the Ca market, since their total market share is approximately 47%.

Company	Market share (%)	Development priority
Pfizer	33.9	--
Bayer	12.8	--
Hoechst	9.0	Third (licensed)
Astra	3.7	Fifth (licensed from Takeda)
BASF	2.7	--
Monsanto (Searle)	2.4	--
Kyowa Hakko	2.2	--
Yamanouchi	2.0	--
Takeda	2.0	Ninth (licensor)
Ciba-Geigy	2.0	

Source: World Review 1999 by IMS Health (The Pharmaceutical Market)

Table 8. "Revenge of success" to market leaders of Ca antagonists, a product not competing with ARBs

Company	ACE Inhibitor Market share (%)	Development priority
Merck Co.	31.0	First
Zeneca	13.4	Fifth (licensed from Tanabe)
Bristol-Meyers Squibb	10.7	Fourth (First as ACE/NEP inhibitors)
Warner-Lambert	6.4	--
Novartis	5.3	Second
Hoechst	3.8	Third
Servier	3.7	Second as ACE/NEP inhibitors
Tanabe	1.9	--
Banyu	1.8	First
Sankyo	1.7	Eighth (licensor to European companies)

Source: World Review 1999 by IMS Health (The Pharmaceutical Market)

Table 9. "Revenge of success" to market leaders of ACEs, a product competing with ARBs

ARB is superior to ACE and differentiated from Ca. From a market viewpoint, ARB competes with ACE directly and replaces the ACE market. The leaders in the ACE market need to develop ARB to keep the current market position because it is obvious that ACE will be replaced once AEB is marketed. In contrast, Ca does not compete with ARB but creates a new market. The leaders in the Ca market do not need to develop ARB to keep their market position in the Ca market, as described above. Surprisingly, the leaders in the Ca market, including Japanese companies, were prohibited from the development of ARB.

Positioning of new product	Superior	Differentiated
Competition with existing product	Direct competing with ACE	Indirect competition or neutral with Ca
Mode of market penetration	Replace old product	Create new market
Attitude to NPD	Enhancing	Inhibitory

Table 10. Two types of new product

This finding demonstrates that a strong existing product inhibits NPD when the product creates a new market, as summarized in Table 10. The most critical reason for the failure of NPD of ARB was the underestimation of the sales forecast, since the sales forecast is basically calculated based on product strength. The company acts to increase the strength of its own product as a market winner in Ca, insisting on the strength of its own product (Monthly Mix, April 1999). This reduces the market value of the new product, creating a new market.

4.3 Case of innovative medical device in the old market

New and innovative medical device technology continues to emerge every year from companies worldwide. Some of these new technologies offer vastly superior capabilities than existing technology, however acceptance of many of these new innovative and superior medical device products often encounter tremendous resistance and neglect by major market players in the marketplace, even when using the outside opinions for evaluation of the product introduction. There appear to be a neglect of emerging market. This section provides the results of a case study of needle-free injection technology from which many valuable findings were derived.

Firms competing in increasingly sophisticated technology markets have encountered a new set of challenges. Responding to customer needs is crucial for survival, while for society as a whole, there are requirements for expanding the reach of technological benefits to larger numbers of individuals. At the firm level, maximizing customer satisfaction by providing an efficient internal manufacturing system and simultaneously securing flexibility corresponding to dynamic and rapid change have become important aspects of any competitive survival strategy. As an inevitable result of too much strengthening of a specific core field, one failure often observed is an inability to quickly move into complementary or different product areas. One survival solution is co-evolution of technology products developed in such a way that external and internal firm circumstances affecting the customer are constantly considered. The question this analysis addresses is, “How do we construct an interface between core and new products in order to simultaneously maximize core competence and yet at the same time remain flexible?”

Institutional elasticity is one mechanism for creating such a trade-off between stability and ongoing new product development. Intriguing in-depth recent case studies on Sears Roebuck, Monsanto, Royal Dutch Shell, the US Army, British Petroleum, Hewlett Packard and Sun Microsystems (Pascale et al., 2000), demonstrate that in business, as in nature, there are no permanent winners. There are just firms that either react to change and evolve, or those that get left behind and become extinct. Equilibrium is a very dangerous position for survival, and innovation usually takes place on the edge of chaos. Self-organization and emergence occur naturally. Organizations are generally more turbulent than directed. Monsanto has successfully remained on the edge of the new business front managing the trade-offs in technology co-evolution. Monsanto is well-known that the company has leading core competence for bio-technology outstandingly different from other biotech firms. Aspartame, artificial sweetner is one of famous product that is produced by comining biotech and chemical technology. Inspite of the former success in food and agricultural NPDs, it could not move beyond its core products. Owing to the failure of NPD in the biotech market, pharmaceutical division of Monsanto is merged by Pharmacia Upjohn in 2000 due to a systemic disconnect between management, technology and market signals. This clearly shows that core competence for technology is not sufficient for successive survival.

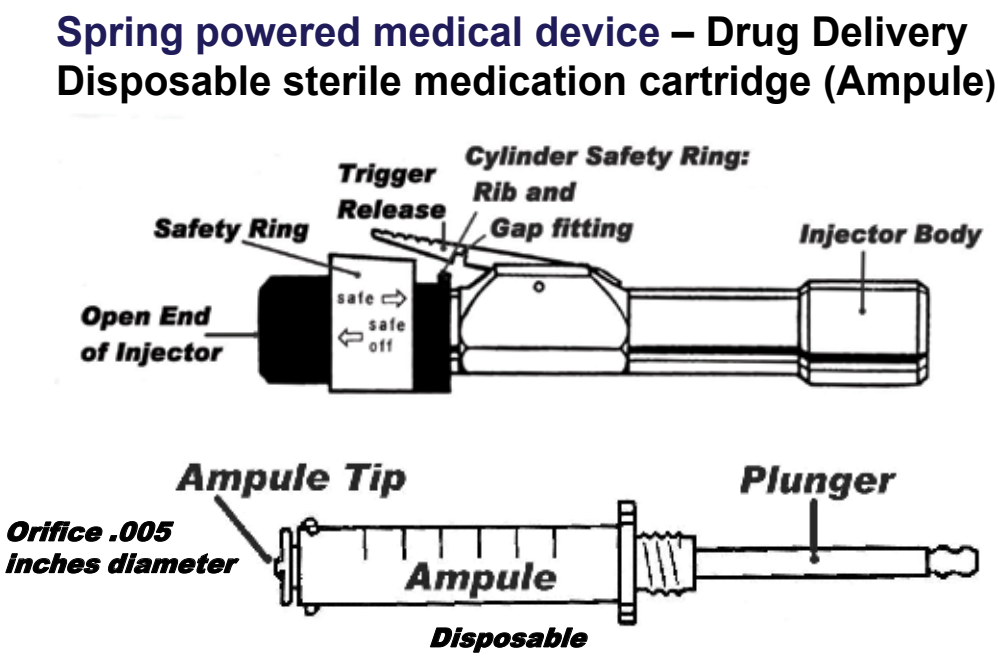


Fig. 3. Construction of needle-free injector device (Inject Co. www.injex.com)

Needle-free drug injection technology is a classic example that reveals the existence of disconnect between technology and market signals, since it is not a new idea. The early and crude beginning of this technology started over 50 years ago, and was later used by the U.S. military to vaccinate military troops in the 1960's. During the 1970's and 1980's others began to conduct extensive R&D to improve, modify and make the technology much more consistent, reliable, pain free and with simple to use ergonomic designs. Today these needle-

free injectors are very small pen-like and very ergonomically designed high tech instruments as shown in Figure 3 and 4.
In this section, INJEX–Equidyne Systems, Inc. is selected since it is considered one of the top two leading companies in the needle-free drug delivery industry and once selected as a partner by a Japanese largest pharmaceutical company and Pharmacia Upjohn before merger with Pfizer (personal communication with those firms). The company is more than fifteen years in business and manufactures and distributes product in over 40 countries worldwide. The company was awarded in 2004 the "medical device technology company of the year" by the well known Frost & Sullivan technology research firm for outstanding and innovative technology as well as excellence in product quality for this industry.

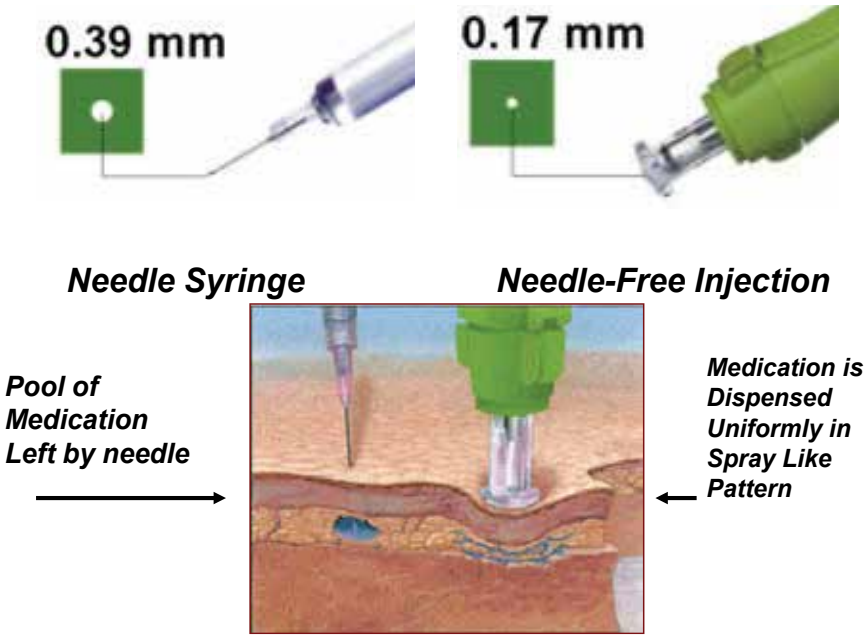


Fig. 4. Mechanism of needle-free syringe for injection through skin (Inject Co. www.injex.com)

After investigation of the issues as well as looking at consumer needs and overall market demand, social issues and other factors, we were able to develop some relevant findings, valuable information and interesting conclusions.
The major pharmaceutical and needle syringe manufacturers have generally overlooked needle-free drug delivery technology. In fact, insulin is injected with tiny-needle syringe by patients themselves every day. Although they claim that they are continuously looking for new and innovative technologies which they can adapt to their current business needs for drug delivery for waccine or pandemic disease like influenza, they have somehow not been able to see the opportunity that needle-free drug injection presents. In this case, each player in the market behave differently since their position to the new product differs each other, as described below in Table 11.
With regard to syringe majors and drug majors, competence is tied to technological core competence, this system works as a strong support for enforcing the original core field of

individual firms. Therefore, it potentially creates inertia which keeps a firm from moving beyond its past. This rigid structure is a cause of "Revenge of Success" for major market players (Takayama, 2002; Takayama & Watanabe, 2002; Takayama, Watanabe & Griffy-Brown, 2002ab; Takayama, 2004). This owes to a rigid scheme adhered to by product evaluation system as internal routine that is related to extensive knowledge of major market players. In case of the injector device majors including the first and second-tiered Japanese majors and the largest major in the world have evaluated this needle-free injector for about one year, conclusively, they decided not to enter the needle-free syringe market. The reason was very simple that syringe business will continue without infringement of evasion of needle-free syringe (interview to those firms).

	Syringe Majors	Drug majors	End User (Patients)
Positioning of new product	Superior in one minor business area	Differentiated	Superior
Competition with existing product	Partially competing	Indirect competition or	Direct competition
Mode of market penetration	Niche or littel market	Create new market	Replace
Attitude to NPD	Neglect	Hesitate	Promote

Table 11. Two types of new products

The scheme of the evaluation routine, which is connected to an open network including outsiders, plays two different roles, depending product positioning of emerging technology in the existing market. (Takayama, Fukushima & Petersen, 2005)

1. In the case of replacing or enhancing an existing product, the open network accelerates the incorporation of new product.
2. On the contrary, when creating a new product market, the open network serves to accelerate the early abandonment of poor in-house NPD and also helps to enhance the rejection of outside strategic alliance.

This practice tends to ensure over-evaluation of a potential new core product evaluation, by the major player in the industry. The open network inevitably brings to any major market player in a high tech market, superior insight and wisdom to analyze and effectively select the correct path for a major success with a new product introduction. This fate is set forth by a cardinal strategy set forth by the major market players, which allows them to compete for top-tier (Takayama, 2004). In conclusion, the fate to win or lose by the market major players is intrinsically highly influenced by "Revenge of Success" though the use of the "Inertia on the open network". In summary, this clearly indicates that open innovation does not work if the market is emerging or newly born from naught.

4.4 Win or loss in the new born market

Following the afore-mentioned discussion, Table 12 and Table 13 summarize mode of competition in some win cases and loss cases of existing majors in the new born markets that are caused by high tech innovation, respectively.

In summary, win cases of existing major in the new born markets are expained by direct competition in the new born market. On the contrary, loss cases of existing major in the new born markets are characterized by indirect or neutral competition of new product to the existing product.

	Existing majors	Mode of competition
Cassette tape recorder	Record player	Direct
CD player	Cassette tape recorder	Direct
Digital camera	Camera film	Direct
DVD	VHS manufacturors	Direct

Table 12. Win cases of existing majors in the new born markets

	Existing majors	Mode of competition
Electrics	Lamp, Mill	Indirect or neutral
Automobile	Carriage	Indirect or neutral
Cassette tape	Viny record	Indirect or neutral
PC	Computor	Indirect or neutral
Amazon	Book store	Indirect or neutral
Net secrities	Secrities	Indirect or neutral
Net retailing	Department store	Indirect or neutral
Net news	Newspaper, TV	Indirect or neutral

Table 13. Loss cases of existing majors in the new born markets

5. Success/failure matrix and replacement of the majors

The importance of the mode of competition of a new product with existing products, regardless if the two are in the same market sector, has so far been amply illustrated so far. The practically deterministic situation may be summarized in a matrix shown in Table 14. An enterprise (major or otherwise) will make serious efforts in developing a new product that may compromise existing core products. In this case, leading companies with a wealth of experience and information associated with those core products are in a highly advantageous position, so that chances for inexperienced newcomers are scarce.

	Direct competition	Indirect competition
Majors	Win	Loss
Newcomers	Loss	Win

Table 14. Win/loss matrix

In contrast, the majors will do nothing if the new product does not compete with their existing products: they lose the opportunity because they are unaware of the market

potential. This mechanism will lead to replacement of the majors if the new product creates its own new market, because it is dominated by newcomers and the majors that would be most familiar with the market are excluded.

	Majors	Newcomers
Regenerative therapy	Pharmaceutical	Fermentation, venture businesses
Gene therapy	Pharmaceutical	Venture businesses
Nano-biotechnology	Pharmaceutical, mechanical	Trading, venture businesses

Table 15. Replacement of the majors in a new market

In case of regenerative medicine, gene therapy and nano-biotech are not regarded as direct competitors to pharmaceuticals. Regeneration of the skin or neurons and regenerative therapy of heat failure are being developed by non-pharmaceutical companies as shown in Table 15. Although these techniques, as well as gene therapy and nano-biotech, will partly replace pharmaceutics in future, drug manufacturers are blind to their market potential. In the near future, pharmacuticals businesses, that insit themselves as representatives of life science business (Pisano, 2006), will lose the future huge market and new born market will be taken over by new comers.

6. Mechanism of success and failure

The mechanism that determines win or loss of a new product development by promoting or hampering corporate R&D works as if by God's hand, as illustrated by the examples presented above. In the highly R&D-intensive bioindustry, market leaders tend to concentrate on their core area, as typified by drug manufacturers. The result is competition for the top tier in a homogeneous environment, which becomes inevitably a rat race (Takayama, 2004). Winners in a market are driven by the "inertia of success", or a desire to continue to succeed, resulting in failure of recognizing the potential of new products outside their own markets (Takayama, Watanabe and Griffy-Brown, 2002a). Management of biobusinesses should take this tendency into account, which has not, however, been well recognized.

Table 16 compares some philosophical aspects of the conventional industries and bioindusty. The former deals with visible physical entities using a common language. Therefore, a competitor's logic is understandable, and individual companies compete with each other in a homogeneous situation. In contrast, the biological world is not entirely logical. Lack of a common language prevents entrepreneurs in this unforeseeable world

	Conventional industries	Bioindustry
Logic	Physical	Biological
Visibility	Visible	Invisible
Predictability	Logical	Ambiguous
Language	Shared	Individual
Competition	Homogeneous	Heterogeneous

Table 16. Philosophical aspects of the conventional industries and bioindusty

from knowing competitors' intentions clearly. Their competition, therefore, tends to occur in a heterogeneous environment. This is reflected in the fact that none of the multifarious biotech products has become a staple commodity.

These differences have not been well recognized. From the viewpoint of conventional industries, the biobusiness is full of tedious processes that are needed before knowing how to manage the unpredictable R&D processes and capricious markets. This accounts largely for the failure of chemical and drug majors in new biotech products.

Corporate goals is in general recognized to be based on the interests of each firm for type of product, social issues, customer preferences and demographics. The general consensus is that improvement of the product acceptance and success is mainly attributed to the company's keen understanding and a strong sense of the customer needs and general market. Surprisingly, the conclusion is exactly the opposite of the general assumptions. Those firms that have a strong interest primarily in product development have failed to keep market share. These firms successfully develop new products on the technology platform. On the contrary to such strong attention to product development R&D and their technology, they may only catch up with the fast movement when a new market was created by other players in the different market.

7. Conclusion

In order to create a firm position in the new born market, strong customer relations seems to be most important influential factor to catch up with the market needs and more customer's satisfaction. Opposition to this general hypothesis, major theories are supported by those that believe that innovative product has been successfully developed by firms who have keen interest in society and human relations including employee relations.

Although these attention definitely seems to contribute to maintain core competency of the firm, most companies prefer to take the easy way to develop new product by connecting to customers and market. Afore-mentioned facts demonstrate that a strong interest in product or corporate profit cannot maintain core competence for product development especially in the case where a new market is created.

What is core competence for creating new market? The strongest core is recognized that human relations and employee relations in the business. The conclusion in this article is that the major firm has a unavoidable fate to deny any new born markers that is differentiated and hence indirectly competed with existing major's products. The law of success or failure of innovation is applicable to not only all high tech driven industry such as the biotech, nanotech, and ICT industry but also new born market that is created by continuous innovation such as new hypertensive pharmaceuticals, specified supplementary foods, down-load music etc. This owes to the neglect of a crucial message from the customers in the new born market like needle-free syringe or regeneration therapy. In spite of major's strong competitive capabilities, the success or loss of each of major players in the new born market is decided by invisible hand of God. How to evade from the fate is the most important implication at the end of this Chapter. All businesses, though, cannot evade from the fate of the law of success or failure. This owes to the neglect of a crucial message from the customers in the new born market like needle-free syringe or regeneration therapy.

8. References

Armbrust, Michael; Fox, Armando; Griffith, Rean; Joseph, Anthony D.; Katz, Randy H.; Konwinski, Andrew; Lee, Gunho; Patterson, David A.; Rabkin, Ariel; Stoica, Ion; and Zaharia, Matei (2009). Above the clouds: A berkeley view of cloud computing. *Technical Report UCB/EECS*-2009-28, EECS Department, University of California, Berkeley

Chesbrough, H (2006). *Open Business Models: How to Thrive in the New Innovation Landscape*, Massachusetts: Harvard Business School Press, ISBN-10: 1422102831, Boston

Christensen, M.C. (1997). *The Innovator's Dilemma.* Harvard Business School Press, Boston

Clerk, K.B.; Fujimoto, T. (1991). Product Development Performance, Harvard Business School Press, Boston

Dosi, Giovanni (1982). Technological paradigms and technological trajectories. *Research Policy* Vol. 2 (3), pp.147–162

Dosi, .Giovanni (1984) *Technical Change and Industrial Transformation*, Macmillan, London

Dosi, Giovanni et al., (1988). *Technical Change and Economic Theory*. Printer Publishers, London

Dosi, Giovanni (2010a). Cambridge Journal of Economics Vol. 34 Issue 1, pp.173-184

Dosi, Giovanni (2010b). Knowledge Accumulation and Industry Evolution: The Case of Pharma-Biotech [Paperback], Mariana Mazzucato and Giovanni Dosi (Eds.) Cambridge University Press, ISBN-10: 052112400X, Cambridge

Fuji Keizai Co. Ltd., 2010 Data book of ethical drugs, mix, 2010/02/24 http://www.mixonline.jp/Article/tabid/55/artid/38622/Default.aspx [6 July 2010]

Gordon, E. Moore (1965). Cramming more components onto integrated circuits" pp.114-117 *Electronics*, Volume 38, Number 8

Gordon E. Moore (1996). Some Personal Perspectives on Research in the Semiconductor Industry, *Engines of Innovation*, Rosenbloom, Richard S., and William J. Spencer (Eds.), pp. 165-174, Harvard Business School Press, ISBN-10: 0875846750, Boston

D. Hamilton. 'Cloud computing' seen as next wave for technology investors. *Financial Post*, 04 June 2008. http://www.financialpost.com/money/story.html?id=562877 [18 July 2008]

von Hippel, E., 1979. *A customer active paradigm for industrial product idea generation,* Baker (ed.)

von Hippel, E., 1980. The user's role in industrial innovation, *Management of Research and Innovation*, Dean, B., Goldhar, J. (Eds.), North Holland, Amsterdam

von Hippel, E. (1982.) Appropriability of innovation benefit as a predictor of the source of innovation. *Research Policy* Vol. 2 (2), 95–116. 623

von Hippel, E. (1988). The Source of Innovation. Oxford University Press, New York

von Hippel, E (2005). *Democratizing Innovation*, The MIT Press ISBN-10: 0262002744

von Hippel, E (2009). *Customers as Innovators: A New Way to Create Value* (HBR OnPoint Enhanced Edition), Harvard Business Review, Boston

Idei, Nobuyuki (2009). Personal communication

Ishi, Junzou. (2010). *President,* Jan. 18,

http://president.jp.reuters.com/article/2010/01/13/A73D8B7A-FF3A-11DE-9D22-0FC03E99CD51.php [6 July 2010]

Freeman, C. (1982). The Economics of Innovation, 2nd ed. Frances Pinter, London

Hammer, M., Champy, J. (1993). Reengineering the Corporation: A Manifesto for Business Revolution. Harper Business, New York

IMS World Review, 1999. The Pharmaceutical Market. IMS Health, London

de Jong, Jeroen P.J. & von Hippel, Eric (2009). Research Policy, Vol. 38 Issue 7, pp1181-1191

JPMA (Japanese Pharmaceutical Manufacturers' Association), (1999a). *Q&A about R&D*, pp. 40–41 (March)

JPMA (1999b). *Hypertensive and its drugs: Q&A 50* (May 1999)

Kokusai Iyakuhin Jouhou (International Drug Information) (1997) Vol. 8(25), pp. 16–19

Meyer, M.H. & Lehnerd, A.P. (1997). *The Power of Product Platforms: Building the Value and Cost Relationship*. The Free Press, New York

Monthly Mix, April 1999. Trend of anti-hypertensives, pp. 36–57

Monthly Mix, September 1999. New class of ATII became the top in hypertensives to get 67.5% of the prescription rate for new patients, pp. 66–68

OECD (1984). Committee for Scientific and Technological Policy, Science, Technology and Competitiveness: *Analytical Report of the Ad Hoc Group. OECD*/STP (84) 26, Paris.

Ohno, T., 1988. *The Toyota Production System*. Productivity Press, Tokyo

Ozawa, Eiji (1970). (in Japanese) Kagaku, Vol. 25, p.854, Iwanami-shoten

Pharma Projects, 1999. V&O Publications, Surrey, UK.

Pisano, G.P., 1997. The Development Factory. Harvard Business School Press, Boston

Pisano, G.P. (2006). *Science Business*, Harvard Business School Publishing, Boston

Rice, MF 2005, ' Information and Communication Technologies and the Global Digital Divide Technology Transfer, Development, and Least Developing Countries' , Comparative Technology Transfer and Society, vol.2, no.2, pp.72-87.

Rothwell, R. et al., 1974. SAPPHO updated. Project SAPPHO, phase 2. *Research Policy* 3 (3), 258–291

Scrip Magazine, February 2000. Leading therapeutics in 1999. PJB Publications, London

Scrip's Yearbook, 1999. PJB Publications, Londn

Takayama, M. (2002). The true reason of failure in new products development (in Japanese), Tokyo Tosho Shuppankai, 2002.

Takayama, Makoto & Chihiro Watanabe (2002). Myth of market needs and technology seeds as a source of product innovation, *Technovation* Vol. 22, pp.353-362

Takayama, Makoto; Chihiro Watanabe & Charla Griffy-Brown (2002a). Remaining Innovative without Sacrificing Stability: An Analysis of Strategies in the Japanese Pharmaceutical Industry that Enable Firms to Overcome Inertia Resulting from Successful Market Penetration of New Product Development, *Technovation* 22, pp.747-759

Takayama, Makoto; Chihiro Watanabe & Charla Griffy-Brown (2002b). The alliance strategy as competitive strategy for successively creative new product development, *Technovation* Vol. 22, 607-614

Takayama, Makoto (2004). Strategy Change from Competition for Top-Tier to Competition for Uniqueness (in Japanese), *Journal of Science Policy and Research Management*, Vol. 19, No 1/2, 58-61.

Takayama, Makoto; Fukushima, Jim & Petersen, Larry (2005). Why Major Needle syringe and Pharmaceutical Manufacturers not yet Accepted the Superior Needle-Free Injection Technology, Even When Using the Kansei Network?, *The Global Business and Technology Association, Reading Books of the Global Business and Technology*, pp.1345-1352

Takayama, Makoto (2005). Win without Fail and Fail without Win in Bio-Management, *Office Automation* Vol. 25, No. 4, pp. 15-21

Takayama, Makoto; Fukushima, Jim & Petersen, L. (2005). Why Major Needle syringe and Pharmaceutical Manufacturers not yet Accepted the Superior Needle -Free Injection Technology, Even When Using the Kansei Network?, *The Global Business and Technology Association, Reading Books of the Global Business and Technology*, pp.1345-1352, The Global Business Association and Technology Association

Takayama, Makoto (2009). Law of success or failure in innovation, *Innovation of Japanese Firm*, Japan Society of Management. (Ed.), ISBN-10: 4805109319, Tokyo

Thrower, P. A. (1999). "Editorial". *Carbon* 37: 1677–1678

Yano Research Institute (2010). *OTC Market Outlook and Strategy 2009*, Tokoyo http://www.yano.co.jp/press/pdf/547.pdf [6 July 2010]

3

Proactive Crisis Management in Global Manufacturing Operations

Yang Liu and Josu Takala
Department of Production, University of Vaasa
PL 700, 65101 Vaasa
Finland

1. Introduction

From an economic perspective, the future has never seemed clear, but high performance businesses have the ability to navigate through uncertainty and emerge ever stronger. How do they do it? The experience and research with the world's most successful companies show that winners follow certain common principles. Companies that come through the strongest actually use economic disruption to improve their competitiveness. This study is to find out how to make it possible.

Future competitiveness of manufacturing operations under dynamic and complex business situations relies on forward-thinking strategies. The objective of this work is to identify and develop the operational competitiveness in a sustainable manner and implement sustainable competitive advantage (SCA), the highly competitive operations strategy by integrating manufacturing and technology strategies with transformational leadership profiles of the decision makers, for managing proactive operations in global turbulent business environments such as the global economic crisis which has badly hit the whole world's economy.

This study is aiming to create methods and tools to analyze the development of operational competitiveness in global context. These include e.g. the following:

- Observation and evaluation of operational strategy excellence and transformational leadership to support decision making processes.
- Scenario analysis of the development of business environments and methods for identifying successful factors of new business concepts with dynamic decision making for optimizing resource allocations by sense & respond methodology and by integrating manufacturing strategy with transformational leadership and technology level to evaluate and benchmark overall operational competitiveness in technology and knowledge intensive business areas.
- Methods and tools for identifying successful factors to develop sustainable operational competitiveness of new business concepts against the highest benchmarks in the world.

2. Literature review

The strategic importance of manufacturing or operations has long been recognised by Skinner (1974). The theoretical reference framework of competitiveness in manufacturing

operations starts from resource-based view of a firm for case study (Wernerfelt 1984; Menguc, Auh & Shih 2007). Companies should typically utilize multi-focused competitive strategies in a holistic way based on their business strategies (Porter 1980). Competitive priorities belong to the first phase of manufacturing strategies, which act as the bridge between business strategy and the manufacturing objectives (Kim & Arnold 1996). Competitive priorities are the crucial decisive variables to manage manufacturing operations in global context and indicate strategies emphasized on developing certain manufacturing capabilities that improve the operational competitiveness. Takala (2002) presents justification of multi-focused manufacturing strategies. Miles & Snow (1978) define four company groups which include prospector, analyzer, defender and reactor. They suggest on the contrary to the three stable groups which are prospector, analyzer and defender, reactor does not lead to a consistent and stable organization and therefore it is advised to change over to one of the other three groups. Based on this theory, Takala et al. (2007b) introduce unique analytical models to evaluate global competitiveness rankings for manufacturing strategies in prospector, analyzer and defender groups according to the company's multi-criteria priority weights of Q (Quality), C (Cost), T (Time/delivery) and F (Flexibility). Such analytical models are used to gain insight into the influences and sensitivities of various parameters and processes on the alteration of manufacturing strategies. In China, the most dynamic market, Liu et al. (2008) first time apply such analytical models to analyze and improve operational competitiveness of one private middle-size manufacturing company by adjusting competitive priorities in manufacturing strategy. Takala et al. (2007a), Si, Takala & Liu (2009), Liu, Si & Takala (2009) and Liu & Takala (2009a; 2010a) compare the operational competitiveness strategies in China and other countries in global context by utilizing same analytical models, in order to analyze different characteristics of manufacturing strategies in different markets and suggest how companies can improve their operational competitiveness. But the adjustment of manufacturing strategy alone is not just enough to improve the overall competitiveness to develop the business under new business situations. Menguc et al. (2007) suggest that improvements of transformational leadership based competencies should lead to marketplace positional advantages through competitive strategies. Therefore manufacturing strategy is one important factor and transformational leadership is another necessary and important factor to improve the overall competitiveness no matter in prosperity or adversity, and can be even more decisive (Bass 1985). Bass & Avolio (1994) provide evidence on the benefits and effectiveness of transformational leadership on leadership and training of leaders. Transformational leaders help their subordinates to learn and develop as individuals, by encouraging and motivating them with versatile repertoire of behavioural and decision making capability (Bass & Avolio 1994; Bass 1997). Takala et al. (2008) introduce unique analytical models to evaluate the level of outcome direction, leadership behaviour and resource allocation of transformational leadership. Tracey, Vonderembse & Lim (1999) suggest that organizations must formulate strategic plans that are consistent with the use of manufacturing technology to be successful in this globally competitive and rapidly changing environment. O'Regan & Ghobadian (2005) suggest that the level of technology deployed will impact on the overall strategic planning process and its main drivers: leadership and organisational culture resulting in differing levels of corporate performance. From these implications, transformational leadership is further extended by adding technology level in this study, which is classified as spearhead technology, core technology, and basic technology, as part of resource allocation. The objective here is to create a holistic

model to integrate manufacturing strategy and transformational leadership with technology level together for more comprehensive evaluation of overall competitiveness to identify and develop the operational competitiveness potential in a sustainable manner.
To validate the created analytical models, the empirical research continues case studies in several countries with deeper insight analysis of overall competitiveness of large and medium size manufacturing enterprises and suggests how to make dynamical adjustments in order to improve operational competitiveness potential to manage in turbulent business situations such as the global financial crisis. The related case studies include benchmarking and development of overall competitiveness of multiple case companies in global context, which emphasize more on proactive operations to improve competitiveness potential in regional and global market during crisis and forecasting the ongoing business in economic upturn after crisis.

3. Research methodologies

3.1 AHP

Analytic Hierarchy Process (AHP) method is a multi-attribute decision instrument that allows considering quantitative, qualitative measures and making tradeoffs (Saaty 1980). The AHP is used in this study to deal with the empirical part, which includes analyzing questionnaires and calculating weights of main criteria and sub-criteria. AHP is aimed at integrating different measures into single overall score for ranking decision alternatives with pair wise comparison of chosen attributes (Rangone 1996). It utilizes pair wise comparisons by interviewing the experts within the whole organization. AHP-based models can comprehensively explore the varying degrees of importance of the indicators and drivers of competitiveness (Sirikrai & Tang 2006). The AHP based instruments e.g. forms and questionnaires have been used in our previous case studies for more than 20 years in successful analysis of case companies and some similar applications of AHP are used in e.g. Zahedi (1989), Rangone (1996), Sun (2004), Banuls & Salmeron (2008), and their validity and reliability are proved. The inconsistency ratio (icr) is calculated to assure the internal validity of pair wise comparison results. Only matrixes with icr value of less than 0.10, and less than 0.30 in smaller groups with competent informants, can be used for reliable decision-making. Otherwise the answers are considered as invalid and will not be used. Further more, some redundant open questions are used in addition to the pair wise comparisons in the AHP questionnaires to add more internal validity to the answers.
The procedures of utilizing the AHP in the case studies are as follow. The first step is to establish the model of hierarchy structure for the goal. In this study, the hierarchy models are constructed for the evaluation of manufacturing strategy by Takala et al. (2007b) and transformational leadership by Takala et al. (2008), which serves as theoretical framework. The second step is the comparison of the alternatives and the criteria. They are pair wise compared with respect to each element of the next higher level. The third step is connecting the comparisons to get the priorities of the alternatives with respect to each criterion and the weights of each criterion with respect to the goal. The local priorities are then multiplied by the weights of the respective criteria. The results are summed up to get the overall priority of each alternative.

3.2 Case study

The empirical research is based on doing numerous case studies of companies from different countries to analyze with existing analytical models and to create new analytical models for

further evaluation, therefore the selection of case companies must be mostly representative, well performed and highly experienced in managing global turbulent business situations. As a result the empirical studies are focused to case companies in the most dynamic market and best performer in crisis management – China, especially the large and medium-sized manufacturing enterprises, and compare their operational performances in global context. The case companies are chosen among the backbone industries of Chinese economy. They cover industries including iron & steel, non-ferrous metal, mining, chemistry, construction, energy, machinery, equipment, research & development, service and logistics. Based on such wide variation of industries and good performance in exercising of strategy and leadership, the chosen case companies are well representative for China in the empirical study.

For side by side comparisons in performance of crisis management, a number of large and median-sized manufacturing case companies in comparable size and similar industries are also chosen from several European countries, including Finland which is known for its highly competitive technologies, Slovakia which is manufacturing base for many European and multinational companies, Spain which is another major European manufacturing centre, and Iceland which is badly hit by the economic crisis. In each country there are around 4 to 5 case companies that are studied. All the case studies in these countries are carried out using exactly the same methodologies as how the case studies are done in China. All case companies are represented with codes which can be neither recognized nor speculated as their real names. The questionnaires for all the case studies are developed based on manufacturing strategy by Takala et al. (2007b) and transformational leadership by Takala et al. (2008).

3.3 Data collection and analysis

The data of case companies in different countries are collected in the same manner, by asking senior managers or directors to answer the questionnaires from different organizations and departments. The interviewees are normally decision makers and middle management groups in the case companies, who have good knowledge about the operations of the case companies, and the number of informants is depended on the size of case company. From same case company the inconsistent results are left out. Firstly, the managers or directors are trained to understand every item of the questionnaires by email, telephone or interview. Secondly, after they finish the questionnaires the answers are analyzed with AHP software. Thirdly, the discussion with managers or directors reveals the results and verifies the validity and reliability of the data further.

For studying the manufacturing strategy, competitiveness priorities are listed in the AHP questionnaires as main criteria consisting of quality, cost, time/delivery, and flexibility. The main criteria are typical items used in evaluating the competitiveness priorities in multi-focused manufacturing strategies (Spina et al. 1996). They are formed based on typical case studies and instruments used in interviews. The sub-criteria involve 19 criterions, such as low defect rate, low cost, fast delivery, broad product line, etc. The weights are statistically measured for further analysis with analytical models (Takala et al. 2007b).

For studying the transformational leadership, leadership profiles are empirically measured with the theoretical frame of reference by AHP questionnaires (Takala et al. 2006). Statistical tests are made to find out the logics in the leadership profiles to increase accuracy in the profiles, and in parallel the analytical models are built by induction and tested statistically to measure leadership skills by leadership indexes from resource utilizations to leadership

behaviours and finally to outcome directions and outcomes. Analytical models are further used to measure the effectiveness of leadership actions within different areas of outcomes and try to find out the correlation between these outcomes and leadership indexes in a forecasting way (Takala et al. 2008).
For studying the technology level, the weights of spearhead technology, core technology, and basic technology are collected by interviewing the expert informants directly (Tuominen et al. 2003).
All the collected answers are further analyzed with analytical models for evaluation of operational competitiveness.

4. Analytical models

In this study, the overall competitiveness is proposed to be evaluated based on two core factors, i.e. manufacturing strategy and transformational leadership. Technology level is proposed to be considered as part of resources of under transformational leadership. Sense & respond model is used to help in dynamic decision making to describe, evaluate, benchmark and optimize lower level resource allocations to meet the performance requirements in all the interest groups inside and outside the organization and in turn to improve higher level strategies.
Existing analytical models of manufacturing strategy and transformational leadership with technology level from Liu & Takala (2009b; 2010b) and sense & respond from Ranta & Takala (2007) are reviewed and examined. These models are integrated to develop as a new holistic model to evaluate and develop overall competitiveness potential.

4.1 Manufacturing strategy

The analytical models for manufacturing strategy are used to calculate the operational competitiveness indexes of companies in the different competitive groups, which are prospector, analyzer and defender (Miles & Snow 1978). According to Takala (2002), the responsiveness, agility and leanness (RAL) holistic model supports the theory of the analytical models using four main criteria, i.e. quality, cost, time and flexibility. The analytical models are developed from our research group based on over 100 case company studies in over 10 countries worldwide, whose industrial branch varies from one to another and company size varies from big to small but they share one thing in common which is that they all compete in a highly dynamic business environment and therefore such analytical models have good transferability.
According to Takala et al. (2007b), the manufacturing strategy index (MSI) is modelled based on the multi-criteria priority weights of Q (Quality), C (Cost), T (Time/delivery) and F (Flexibility), as function $MSI = f_{MSI}(Q, C, T, F)$.
The equations to calculate normalized weights of core factors are as follow.

$$Q' = \frac{Q}{Q + C + T} \tag{1}$$

$$C' = \frac{C}{Q + C + T} \tag{2}$$

$$T' = \frac{T}{Q + C + T} \tag{3}$$

$$F' = \frac{F}{Q + C + T + F} \tag{4}$$

Q = Quality; C = Cost; T = Time/delivery; F = Flexibility

The analytical models to calculate the manufacturing strategy indexes of operational competitiveness in each group (Prospector, Analyzer, Defender) are respectively as follow.

$$MSI_P = 1 - (1 - Q'^{1/3}) \cdot (1 - 0.9 \cdot T') \cdot (1 - 0.9 \cdot C') \cdot F'^{1/3} \tag{5}$$

$$MSI_A = 1 - (1 - F') \cdot \left(abs \begin{Bmatrix} (0.95 \cdot Q' - 0.285) \cdot (0.95 \cdot T' - 0.285) \cdot \\ (0.95 \cdot C' - 0.285) \end{Bmatrix} \right)^{1/3} \tag{6}$$

$$MSI_D = 1 - (1 - C'^{1/3}) \cdot (1 - 0.9 \cdot T') \cdot (1 - 0.9 \cdot Q') \cdot F'^{1/3} \tag{7}$$

4.2 Transformational leadership with technology level

The theoretical frame of the analytical models is based on theory of transformational leadership (Bass 1997). A holistic but very simple model of a human being from resource allocations to behaviour and finally to outcome directions and outcomes has been built basing on psychic, social, functional, organizational and structural factors and put together according to the sand cone model and participation objectives in leadership of an organization (Takala et al. 2006). A modified sand cone model by integrating technology level into part of resource is proposed in Liu & Takala (2010b), based on which the new analytical models are developed. Sand cone model from operations management literature Ferdows & De Meyer (1990) present a model of cumulative layers of manufacturing performance dimensions. The model implies an idea that companies need to develop their performance in certain stages, in order to achieve higher levels of competitive performance. The prescriptive order of mutually supportive and enabling success factors is to proceed from quality, to delivery performance, then flexibility and finally to cost effectiveness. In this manner, the often-competitive dimensions of performance need to be viewed as a whole, to think about performance and capabilities on a longer-term basis. The conceptual model with sand cone has similar basic ideas as the model of deep leadership (Nissinen 2001) in which the potential in professional skills and resources is transformed to outcomes of activities with the help and support of leadership process and behaviour.

Technology is understood as know-how of human competence, a relevant part of resource based strategy, including all types of assets and resources, or strategic networking for collaborations by using partnerships (Braun 1998; Takala 1997). The technology level, which is categorized as spearhead technology (SH), core technology (CR), and basic technology (BS), are defined as follow.

SH: Technologies that are more orientated for the future.

CR: Core competitive technologies that are in use for today.

BS: Technologies that are commonly used everywhere and can be outsourced or purchased from other companies.

Based on analytical models for transformational leadership proposed by Takala et al. (2008), the analytical models are further developed by integrating technology into resource for the evaluations of leadership indexes and its outcomes of transformational leadership. These models are outcome direction index (OI) by balancing the directions, leadership behaviour index (LI) by measuring deep leadership, the maximum of passive and/or controlling leadership and the utilization of the cornerstones of deep leadership in different ways, and resource allocation index (RI) by balancing utilization of human resources. Outcome index (OI) is based on weights of factors i.e. extra effort (EE), satisfaction (SA), effectiveness (EF), therefore OI is modelled as function $OI = f_{OI}(EE, SA, EF)$. Leadership index (LI) is based on weight of factors i.e. deep leadership (DL), passive leadership (PL), controlling leadership (CL) and individualized consideration (IC), inspirational motivation (IM), intellectual stimulation (IS), building trust and confidence (BT), therefore LI is modelled as function $LI = f_{LI}(DL, PL, CL, IC, IM, IS, BT)$. Resource index (RI) is based on weights of factors i.e. people/technology/know-how (PT), processes (PC), information systems (IT), organizations of groups/teams (OR) and technology level index (TI), where TI is based on weights of factors i.e.: spearhead technology (SH), core technology (CR), and basic technology (BS), therefore TI is modelled as function $TI = f_{TI}(SH, CR, BS)$ and RI is modelled as function $RI = f_{RI}(PT, PC, IT, OR, TI)$. The total leadership index (TLI) is still modelled as function $TLI = f_{TLI}(OI, LI, RI)$ as in previous studies, however, the difference of the new definition of TLI is that TI has been considered to be integrated into transformational leadership as a special part of RI in leadership.

The modelling of technology level is different than other variables because of its particularity in transformational leadership. Here a brand new idea to model the effect of technology level index to resource index is proposed. According to the principles how resource index has been built, the effects are defined as follow.

A. The excessive know-how, meaning that caused by not the right technology belongs directly as an extra weight to the warehouse of know-how (PT), and lowers weights in PC, IT or OR, lowering in both cases the resource index RI in a linear manner.

B. The right technology, meaning that fitting to the manufacturing stages increases PC, IT or OR, and decreases the know-how (PT) warehouse that caused by not the right technology, and increases in both cases the resource index RI in a linear manner.

Definitions A and B with the expert opinions from the case companies and equation for modelling RI are used for the analysis. The weights of SH/CR/BS are collected by interviewing the experts especially how significant or how much effect they are or have to be for PT and min(PC, IT, OR) and then the effects of how TI affects RI is analyzed.

Assuming that followed by previous business situation there are new business situations of an economic downturn and then an economic upturn, companies need to deal with the crisis and then recover from the crisis. One example to analyze how TI might affect RI in three phases of different business situations, i.e. before crisis, during crisis, after crisis, is presented in Table 1.

The optimal weights of SH, SR, and BS under different stages of crisis are listed in Table 1. These optimal values are obtained theoretically from the chosen competitor and market benchmark with some tolerance. Then the case company data are compared with the optimal values to get the differences for calculating TI. TI is defined to reflect how good the technology level allocation is by using 1 minus the worst deviation from the optimal weights of technology levels. The higher value of TI directly decrease PT caused by using

	Before crisis (BC)	During crisis (DC)	After crisis (AC)
SH	High, factor 2…, ≥60%	Lower, factor about 1, 20%~30%	High, factor 1.5…2, 45%~70%
CR	Low, factor 1…, ≥30%	Higher, factor about 2, 40%~60%	Lower, factor …1, ≤35%
BS	About 0, ≤10%	Low, factor 0.5…1, 10%~30%	About 0, ≤10%
RI	=RI(BC), with PT low and min(PC, IT, OR) high	=1.2…2×RI(DC), with PT higher and min(PC, IT, OR) lower	=1.05…1.2×RI(AC), with PT high and min(PC, IT, OR) lower
TLI	=TLI(BC)	=1.2…2×TLI(DC)	=1.05…1.2×TLI(AC)

Table 1. How TI affects RI under different business situations

not the right technology and increase min(PC, IT, OR), therefore increases RI eventually. Based on such idea, TI is modelled.

The analytical models for evaluation of leadership are as follow.

Outcome index (OI):

According to Liu & Takala (2009b: 13), different categories of outcome indexes all lead to nearly same total dealership indexes, therefore this empirical research uses OI model without classification:

$$OI = 1 - \max\left\{\left|\frac{1}{3} - EE\right|, \left|\frac{1}{3} - SA\right|, \left|\frac{1}{3} - EF\right|\right\} \tag{8}$$

Categorized OI models (Takala, Kukkola & Pennanen 2008; 2009) are preliminary and will be explored more in future research.

The OI model for prospector group:

$$OI_P = 1 - (1 - EE^{1/3}) \cdot (1 - EF) \cdot (1 - SA) \cdot Std\{EE, SA, EF\}^{1/3} \tag{9}$$

where EE ≥ 0.43 and EF+SA ≤ 0.57

The OI model for analyzer group:

$$OI_A = 1 - (1 - SA^{1/3}) \cdot (1 - Std\{EE, SA, EF\}^{1/3}) \tag{10}$$

where SA ≥ 0.43 and EE+EF ≤ 0.57

The OI model for defender group:

$$OI_D = 1 - (1 - EF^{1/3}) \cdot (1 - EE) \cdot (1 - SA) \cdot Std\{EE, SA, EF\}^{1/3} \tag{11}$$

where EF ≥ 0.43 and EE+SA ≤ 0.57

The OI model for reactor group:

$$OI_R = (OI_P + OI_A + OI_D)/3 \tag{12}$$

where EE < 0.43 and SA < 0.43 and EF < 0.43

EE = extra effort; SA = satisfaction; EF = effectiveness

Leadership index (LI):

$$OI_R = (OI_P + OI_A + OI_D)/3 \tag{13}$$

DL = deep leadership; PL = passive leadership; CL = controlling leadership
IC = individualized consideration; IM = inspirational motivation;
IS = intellectual stimulation; BT = building trust and confidence
Resource index (RI) integrating with Technology index (TI):

$$RI = (1 - PT \cdot (1 - TI)) \cdot (3 \cdot \min\{PC, IT, OR\} \cdot TI) \tag{14}$$

PT = people, technology and know-how; PC = processes;
IT = information systems; OR = organizations (groups, teams)

$$TI = 1 - \max\left\{\left|SH_{optimal} - SH\right|, \left|CR_{optimal} - CR\right|, \left|BS_{optimal} - BS\right|\right\} \tag{15}$$

SH=Spearhead; CR=Core; BS=Basic
Combined total leadership index (TLI):

$$TLI = OI \cdot LI \cdot RI \tag{16}$$

4.3 Overall competitiveness

The overall competitiveness index (OCI) is proposed to be modelled as function:

$$OCI = f_{OCI}(f_{MSI}, f_{TLI}) = f_{MSI} \cdot f_{TLI} = MSI \cdot TLI \tag{17}$$

According to Liu & Takala (2009b: 14), in some cases the OCI can be modelled as reduced function:

$$OCI = f_{OCI}(f_{MSI}, f_{TLI}) = f_{MSI} \cdot f_{TLI} = MSI \cdot OI \cdot TI \tag{18}$$

This is because that the OI of transformational leadership is the key factor to direct the strategic goal of manufacturing strategy and MSI is the driving force of the company, taking the effects of TI into account in which TI are evaluated as approximately constant factors before crisis, during crisis and after crisis. In such cases, OI is more decisive to overall competitiveness but other factors like LI, RI, and TI can be influenced also by government macro control, etc.

5. Empirical research

In this chapter, a complete example by studying the development of operational competitiveness potential of case companies in China, Finland, Slovakia, Spain and Iceland is presented to illustrate the applicability of implementing SCA for proactive operations to manage in economic crisis situation.

5.1 Overview of analysis process

The collected answers from questionnaires are processed and analyzed step by step for evaluation of operational competitiveness and development of operational competitiveness potential. Fig.1 shows the complete process of the empirical research from questionnaires to conclusions illustrated with a flowchart.

5.2 Data processing and analysis

AHP analysis of raw data is the first step of the process. Raw data from answers of questionnaires are processed with AHP software Expert Choice, to convert qualitative criteria to quantitative values. During this step, inconsistency ratios are checked to ensure the internal validity of the answers. Also the results are compared with answers to open questions for added internal validity to the answers. Different business situations, e.g. before crisis, during crisis, after crisis are processed respectively.

Analytical evaluation of MSI and TLI is the second step of the process. The results from AHP are further processed with analytical models introduced in section 4. All the analytical models are programmed with Matlab code for the ease of processing data.

The rankings of MSI are obtained from our global manufacturing strategy database which composes of MSI case studies of over 100 case companies in over 10 countries worldwide. The most competitive case companies in prospector group are from Iceland, China and Finland; in analyzer group are from China, Spain and Finland; in defender group are from Spain, Iceland and China, and the most powerful transformational leadership of case company leaders are from China and Finland.

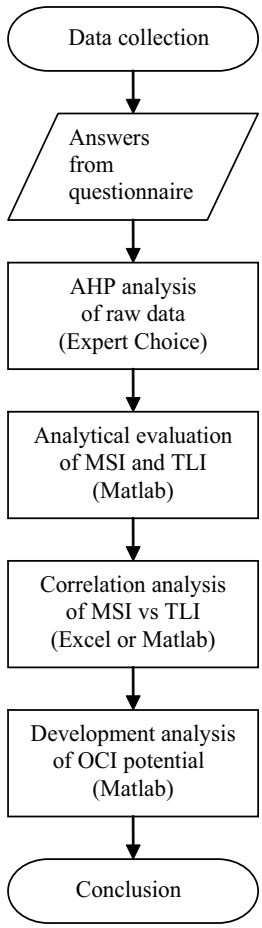


Fig. 1. Flowchart for empirical research

Correlation analysis of MSI vs TLI is the third step of the process. In an organization, TLI is considered as driving force and MSI is considered as outcome, therefore it's meaningful to find the correlation of MSI vs TLI. The results of the case companies, in this example divided by countries, or smaller units such as regions or industries or companies, are plotted with Excel or Matlab to show the correlations of MSI in different groups (prospector, analyzer and defender) versus TLI. The smaller the divided units, the more accurate are the results. In each divided unit to be analyzed, at least 3 answers for each competitive group are required, which make it possible to provide sufficient information for measuring the significance of regression in order to analyze the OCI potential, and more answers reflect the reality better.

Fig. 2, Fig. 3, Fig. 4, Fig. 5, Fig. 6 respectively plot the correlations between MSI and TLI of case companies in China, Finland, Slovakia, Spain and Iceland. It can be seen that the slopes of MSI vs TLI in different groups are quite different. Typically, the group which has the highest slope and the highest significance of the regression measured by R-square is considered to be the most competitive group in the divided unit under that particular business situation.

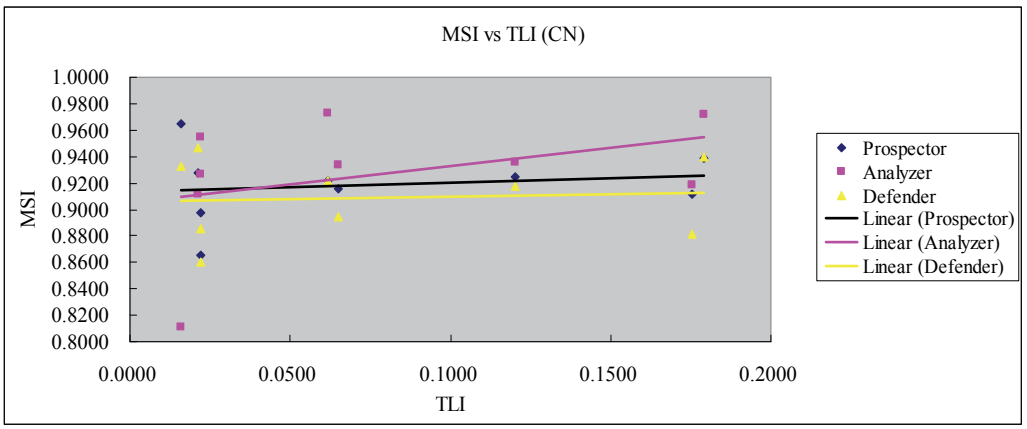


Fig. 2. MSI vs TLI of case companies in China

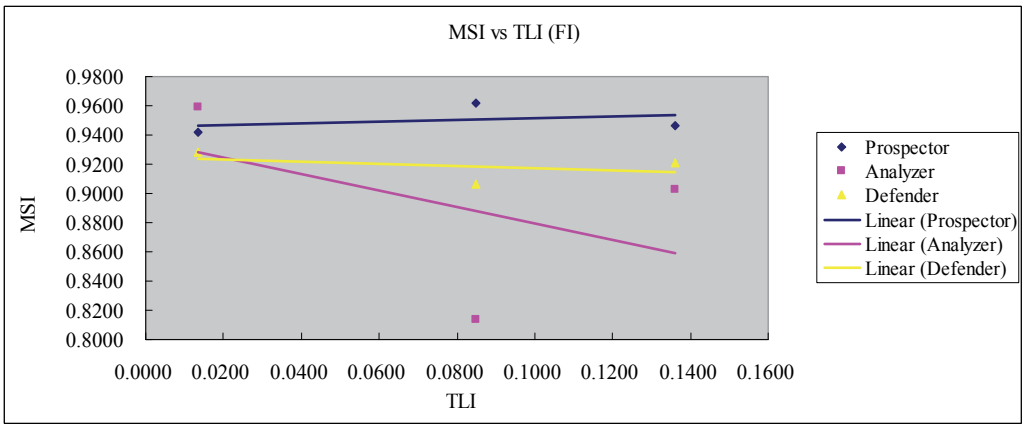


Fig. 3. MSI vs TLI of case companies in Finland

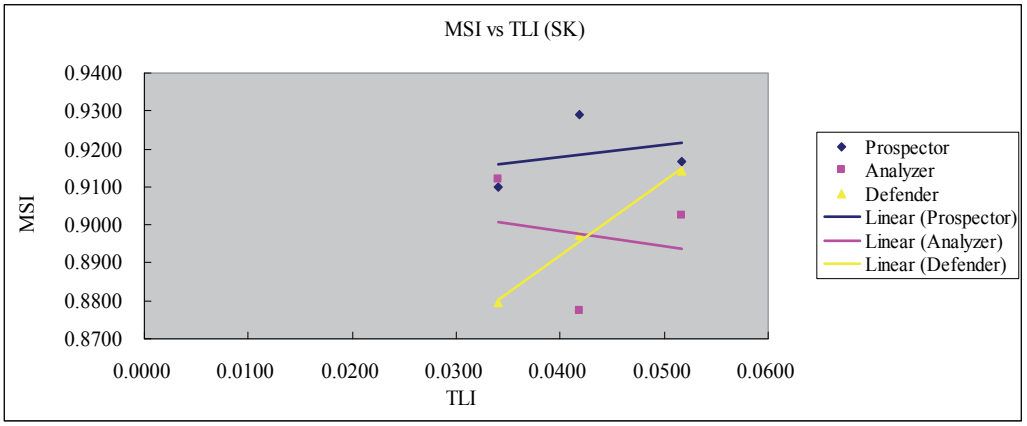


Fig. 4. MSI vs TLI of case companies in Slovakia

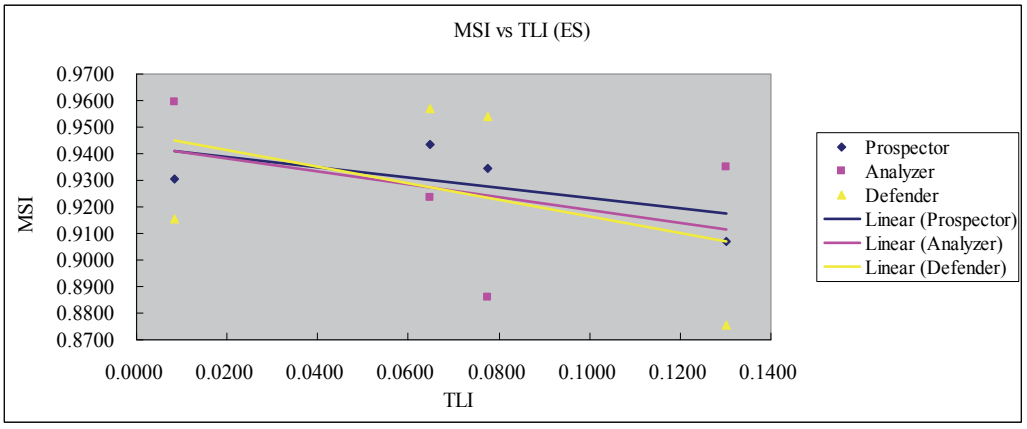


Fig. 5. MSI vs TLI of case companies in Spain

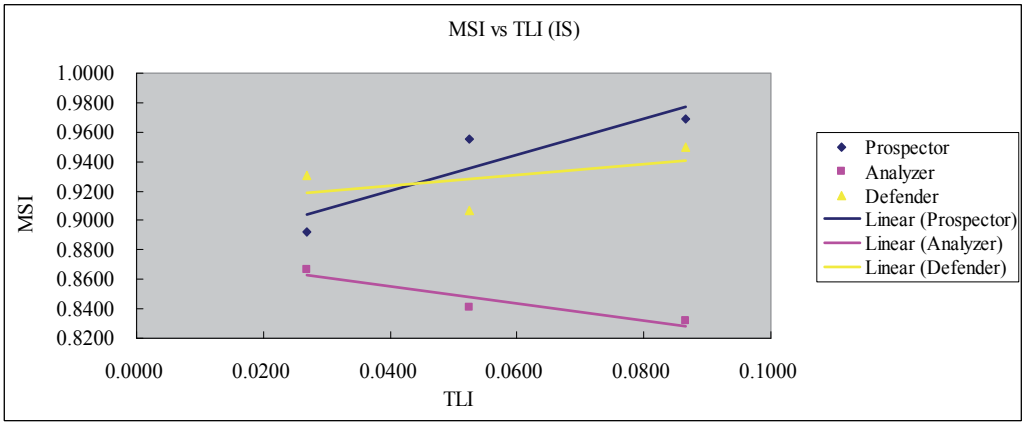


Fig. 6. MSI vs TLI of case companies in Iceland

Development analysis of OCI potential is the fourth step of the process and the most important one. To develop the operational competitiveness potential with the most competitive group in the particular unit, the idea is to break the links between each leader's TLI and the corresponding MSI so that the leader's full potential TLI can be utilized to drive the best possible MSI and in turn to obtain the highest possible OCI potential. This gives "what if" assumptions that leaders are believed to be able to generate better operational competitive performance if they are switched to more suitable positions.
In the particular unit the MSI from most competitive group and corresponding TLI are independently sorted from low to high, and plotted respectively against number of samples. An example of case companies in China with sorted MSI in most competitive analyzer group and sorted TLI form the two linear regression functions as shown in Fig. 7 with

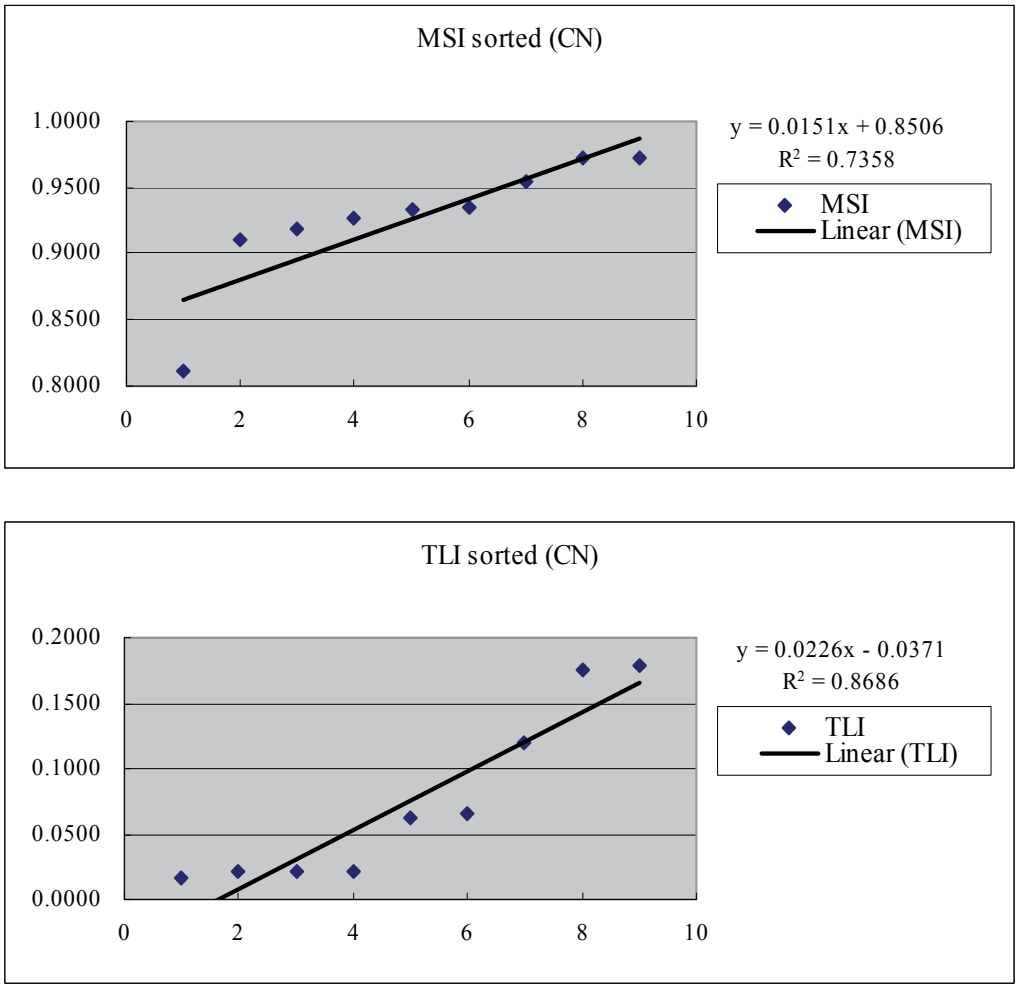


Fig. 7. Linear regression functions of sorted MSI and TLI

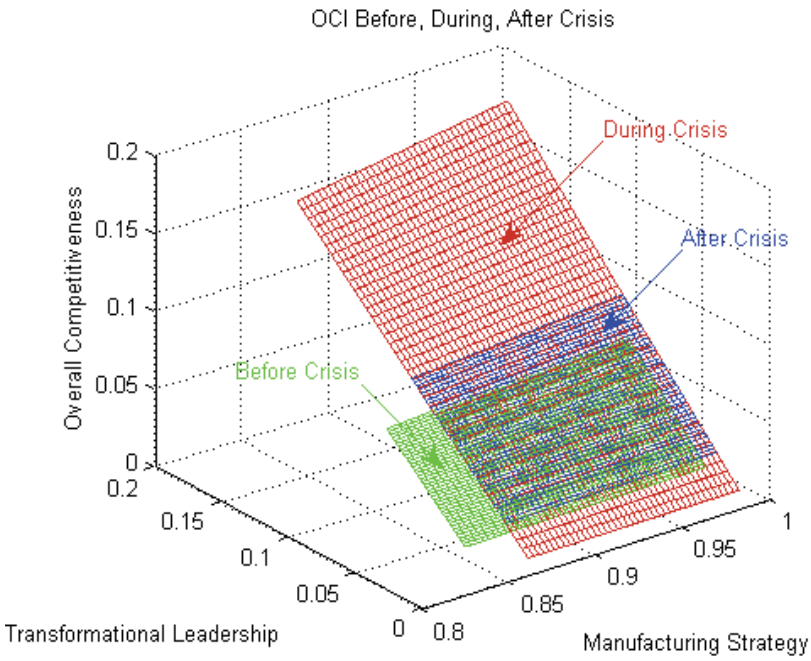


Fig. 8. OCI case comparisons before, during, after crisis

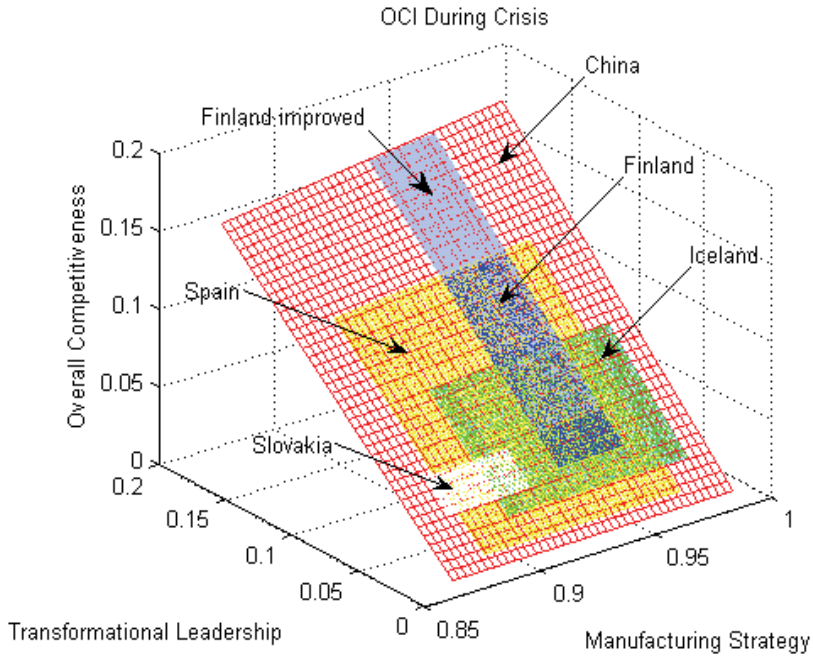


Fig. 9. Improved OCI during crisis with case comparisons

relatively high R-squares can be obtained. The product of these two functions is plotted in Matlab with 3-dimensional mesh function to show the potential region where the OCI can be developed. The plots of OCI potential regions can be used both for horizontal comparisons e.g. to compare the same unit under different business situations and for vertical comparisons e.g. to compare different units under the same business situation. Fig. 8 shows an example of horizontal comparisons which compares OCI before, during and after crisis for the same unit. Fig. 9 shows an example of vertical comparisons which compares OCI from different units during crisis. Through sense & respond model to optimize resource allocations for case companies in Finland, the improved OCI potential regions can be forecasted.

OCI potential analysis and such comparisons can be very helpful to study the effects of dynamic decisions on operational competitiveness under different business situations and develop the competitiveness potential further.

5.3 Findings

The evaluation results of MSI show that case companies from different countries have demonstrated by using different strategies to deal with economic crisis. In prospector group, Icelandic case companies have shown strong competitiveness despite of the seriously threatened economy from Icelandic banking crisis, and the evaluation results indicate that prospector strategy has successfully maintained their best competitiveness to survive during crisis. In analyzer group, Chinese case companies mostly have maintained or have changed to analyzer as the most competitive group and have shown strong competitiveness during crisis. This can be explained by the fact that during crisis the significant decrease in market demand has driven a strict control over costs both in production and administration. Also externally, Chinese government plays a key role at the macroeconomic level which regulates the domestic market more than other case countries (Si et al. 2010). In defender group, Spanish case companies have shown strong competitiveness which reflects realistically that cost effectiveness is their competitive advantage to sustain market shares during crisis.

The evaluation results of TLI indicate that Chinese leaders have demonstrated the most powerful transformational leadership, and this result is consistent with the result from manufacturing strategy. By reviewing evaluation results of MSI, Chinese case companies have shown relatively strong competitiveness in all three different groups. This can be explained by the fact that Chinese government regulation applied to different industries has pointed leaders the clear direction of development, so that they have more clear vision for taking actively courage to the challenge of crisis and making the right adjustments in dealing with the crisis. On the contrary, also some leaders have been left behind, which implies they are passively waiting for the government solutions and trying adjustments in the wrong way because of lacking experience.

From the correlation analysis it can be seen that in different case countries or even under different business situations the slopes of MSI vs TLI may be positive or negative. Especially in Fig. 5 the cases in Spain where MSI vs TLI in all groups have negative slopes, which implies good transformational leadership does not necessarily lead to strong competitiveness in manufacturing strategy. Only exception is the cases in China as shown in

Fig. 2 where MSI vs TLI in all groups have positive slopes, which implies during crisis leadership is really motivated and plays a key role in dealing with the crisis. The slope of MSI vs TLI in analyzer group is highest, which proves transformational leadership makes manufacturing strategy more competitive in analyzer group during crisis. Fig. 3 and Fig. 6 demonstrate how TLI affects MSI for case companies in Finland and Iceland. The slopes of MSI vs TLI indicate they have highest competitiveness in prospector group. These results are corresponding to the fact that most of Finnish firms are advanced in technology as well as good vision in future product development, and Icelandic firms have survived during crisis by aggressively searching for new market and profit from products innovations despite of the serious impact of economic downturn on its economy. Fig. 4 implies that case companies in Slovakia have the highest competitiveness in defender group, which can be explained in practice that Slovakia firms are very cost efficient, and during crisis they use low cost strategy to obtain the market, however the results are not very promising. A general finding is that leaders in China do show active adjustments during crisis, whereas leaders in other countries seem more conservative and that even limit further the competitiveness of manufacturing and technology strategies, causing negative slopes of MSI vs TLI. It is assumed that in China leaders are adventurous to make more competitive decisions since they have strong support from government so that they don't worry too much in taking aggressive decisions. Another significant finding in MSI vs TLI is that typically in analyzer group the plots are scattered which results very low R-square values, whereas in prospector and defender groups the plots usually result relatively much higher R-square values. The causes of such phenomenon will be dealt in future research.

The 3-dimentional plots in Fig. 8 and Fig. 9 show the competitiveness potential of case companies under different business situations and in different countries where the OCI can be developed. It can be seen that transformational leadership has more significant effect than manufacturing strategy to improve overall competitiveness potential. Through such proactive operations to develop sustainable competitive advantage, the forecasted OCI after crisis shows continuous improvement over previous OCI before crisis in Fig. 8 and the forecasted OCI improvement is significant over the previous one in Fig. 9, where the research goal of this work is reached.

5.4 Summary

Compared to previous research results which have been conducted before crisis, such comparative studies to place a number of case studies longitudinally to examine the impact of economic crisis is a unique opportunity to find the solution of how to overcome the crisis and recover after the crisis.

To conclude this empirical research, China shows strong potential in developing overall operational competitiveness compared to other countries, which might explain China's leading role in dealing with global economic crisis from operations point of view. This can be further proved by official statistics. According to The World Bank (2010)'s latest China Quarterly Update released in March 2010, China's economy grew 8.7 percent in 2009 and the growth momentum continued in the first months of 2010 in spite of the global recession. The adjustments in manufacturing strategy and transformation leadership by implementing

SCA through fast strategy are proved to be effective and successful to manage the crisis and keep the high growth of Chinese economy. The experience learnt from this research can thus become a model for crisis management studies globally.

6. Conclusion

This work studies the evaluation and development of overall operational competitiveness in global context using analytical models, which is a novel concept by integrating the evaluation of manufacturing strategy and transformational leadership with technology level together, and through sense & respond proactive operations to improve the competitiveness potential in order to manage turbulent business situations. The empirical research is focused to numerous case studies of companies in China and several European countries to compare their overall competitiveness in global context and conclude the experience of managing the economic crisis, with the purpose in finding solutions to manage turbulent business situations. The influence of "China effect" and global economic crisis are brought together to study how such will impact the operational competitiveness of companies on top of their previous manufacturing strategy and transformational leadership before crisis, and how they will react during crisis to adjust their current manufacturing strategy and transformational leadership to manage the crisis, and even to predict after crisis how they will minimize the negative impacts from the crisis to sustain and develop their optimal operational competitiveness further. The competitiveness in manufacturing operations is evaluated in terms of overall competitiveness performance by integrating the core factors i.e. manufacturing strategy and transformational leadership with technology level, into conceptual analytical models, and through sense & respond to optimize resource allocations to help in dynamic decisions in adjusting strategies and transforming leadership in order to improve the competitiveness potential in a sustainable manner. Implementing such strategic adjustments and transformational capabilities are proposed as the unique SCA for managing in global turbulent business environments.

7. References

Banuls, V. A. & Salmeron, J. L. (2008). Foresighting key areas in the Information Technology industry. *Technovation* 28: 3, 103-111.

Bass, B. M. (1985). *Leadership and Performance beyond Expectations*. New York: Free Press.

Bass, B. M. & Avolio, B. J. (1994). *Improving Organizational Effectiveness through Transformational Leadership*. Thousand Oaks: Sage.

Bass, B. M. (1997). Does the transactional-transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist* 52: 2, 130-139.

Braun, E. (1998). *Technology in Context: Technology Assessment for Managers*. London: Routledge.

Ferdows, K. & De Meyer, A. (1990). Lasting improvements in manufacturing performance: In search of a new theory. *Journal of Operations Management* 9: 2, 168-184.

Kim, J. S. & Arnold, P. (1996). Operationalizing manufacturing strategy: An exploratory study of constructs and linkage. *International Journal of Operations & Production Management* 16: 12, 45-73.

Liu, Y., Li, Y., Takala, J., Kamdee, T. & Toshev, R. (2008). Improve company's operative competitiveness using analytical models. *Proceedings of the 17th International Conference on Management of Technology – IAMOT 2008*. Dubai: International Association for Management of Technology.

Liu, Y., Si, S. & Takala, J. (2009). Comparing operational competitiveness strategies in China and Finland. *Proceedings of the 18th International Conference on Management of Technology – IAMOT 2009*. Orlando: International Association for Management of Technology.

Liu, Y. & Takala, J. (2009a). Crisis management of Chinese state-owned manufacturing enterprises in global context. *Proceedings of Management International Conference – MIC 2009*. Sousse.

Liu, Y. & Takala, J. (2009b). Modelling and evaluation of operational competitiveness of manufacturing enterprises. *Quality Innovation Prosperity* XIII: 2, 1-19.

Liu, Y. & Takala, J. (2010a). Evaluation of global operational competitiveness for crisis management. *Proceedings of the 19th International Conference on Management of Technology – IAMOT 2010*. Cairo: International Association for Management of Technology.

Liu, Y. & Takala, J. (2010b). Competitiveness development of Chinese manufacturing enterprises in global context for crisis management. *International Journal of Management and Enterprise Development* X: Y, 000-000. (forthcoming)

Menguc, B., Auh, S. & Shih, E. (2007). Transformational leadership and market orientation: Implications for the implementation of competitive strategies and business unit performance. *Journal of Business Research* 60: 4, 314-321.

Miles, R. E. & Snow, C. C. (1978). *Organizational Strategy, Structure, and Process*. New York: McGraw-Hill.

Nissinen, V. (2001). *Military leadership, a critical constructivist approach to conceptualizing, modelling and measuring military leadership in the Finnish defence forces*. Finnish National Defence University. Department of Leadership and Management. Dissertation.

O'Regan, N. & Ghobadian, A. (2005). Strategic planning – a comparison of high and low technology manufacturing small firms. *Technovation* 25: 10, 1107-1117.

Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.

Rangone, A. (1996). An analytical hierarchy process framework for comparing the overall performance of manufacturing departments. *International Journal of Operations & Production Management* 16: 8, 104-119.

Ranta, J. M. & Takala, J. (2007). A holistic method for finding out critical features of industry maintenance services. *International Journal of Services and Standards* 3: 3, 312-325.

Saaty, T. L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York: McGraw-Hill.

Si, S., Takala, J. & Liu, Y. (2009). Competitiveness of Chinese high-tech manufacturing companies in global context. *Industrial Management & Data System* 109: 3, 404-421.

Si, S., Liu, Y., Takala, J. & Sun, S. (2010). Benchmarking and developing the operational competitiveness of Chinese state-owned manufacturing enterprises in a global context. *International Journal of Innovation and Learning* 7: 2, 202–222.

Sirikrai, S. B. & Tang, J. C. S. (2006). Industrial competitiveness analysis: Using the analytic hierarchy process. *The Journal of High Technology Management Research* 17:1, 71-83.

Skinner, W. (1974). The focused factory. *Harvard Business Review* May-June, 113-121.

Spina, G., Bartezzaghi, E., Bert, A., Cagliano, R., Draaijer, D. & Boer, H. (1996). Strategically flexible production: the multi-focused manufacturing paradigm. *International Journal of Operations & Production Management* 16: 11, 20-41.

Sun, S. (2004). Assessing joint maintenance shops in the Taiwanese Army using data envelopment analysis. *Journal of Operations Management* 22: 3, 233-245.

Takala, J. (1997). Developing new competitive strategies for high performance organizations from empirical case studies on relationship between technology management and total quality management. *Proceedings of International Conference on Productivity and Quality Research – ICPQR 1997*. Houston.

Takala, J. (2002). Analyzing and synthesizing multi-focused manufacturing strategies by analytical hierarchy process. *International Journal of Manufacturing Technology and Management* 4: 5, 345-355.

Takala, J., Leskinen, J., Sivusuo, H., Hirvelä, J. & Kekäle, T. (2006). The sand cone model: illustrating multi-focused strategies. *Management Decision* 44: 3, 335-345.

Takala, J., Hirvela, J., Liu,Y. & Malindzak, D. (2007a). Global manufacturing strategies require "dynamic engineers"? Case study in Finnish industries. *Industrial Management & Data System* 107: 3, 328-344.

Takala, J., Kamdee, T., Hirvela, J. & Kyllonen, S. (2007b). Analytic calculation of global operative competitiveness. *Proceedings of the 16th International Conference on Management of Technology – IAMOT 2007*. Orlando: International Association for Management of Technology.

Takala, J., Pennanen, J., Hiippala, P., Maunuksela, A. & Kilpiö, O. (2008). Decision maker's outcome as a function of transformational leadership. *Proceedings of the 17th International Conference on Management of Technology – IAMOT 2008*. Dubai: International Association for Management of Technology.

The World Bank (2010). *China Quarterly Update – March 2010* [Web document]. Beijing: World Bank Office [Cited on 11 May 2010]. Available at: http://siteresources.worldbank.org/CHINAEXTN/Resources/318949-1268688634523/CQU_march2010.pdf

Tracey, M., Vonderembse, M. A. & Lim, J. S. (1999). Manufacturing technology and strategy formulation: keys to enhancing competitiveness and improving performance. *Journal of Operations Management* 17: 4, 411-428.

Tuominen, T., Rinta-Knuuttila, A., Takala, J. & Kekäle, T. (2003). Technology survey: logistics and automation branch of materials handling industry. *Proceedings of the 2nd International Conference on Logistics & Transport – LOADO 2003*. High Tatras.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategy Management Journal* 5: 2, 170-180.

Zahedi, F. (1989). Quantitative evaluation of micro versus larger database products. *Computers & Operations Research* 16: 6, 513-532.

4

Signals for Emerging Technologies in Paper and Packaging Industry

Karvonen Matti and Kässi Tuomo
Lappeenranta University of Technology
Finland

1. Introduction

The paper industry is undergoing significant changes in its business environment as both media and the packaging industries are constantly striving for inexpensive methods to add new functionalities in their products and to develop their processes. Other industries, such as the ICT, electronics and food industry look at printing as an economical method for mass production which also creates new applications and opportunities. The development has lead to the convergence or fusion of technologies. In previous research the phenomenon of convergence has received particular attention within the information and communication technologies (ICT) (e.g. Duysteers & Hagedoorn, 1998; Lei, 2000; Stieglitz, 2003; Wirtz, 2001). Despite the fact many industries are facing trends of convergence, the phenomenon has remained largely unexplored in the academic management field. Blurring, or even disappearance, of industry boundaries, overlapping technologies and markets are used to describe this phenomenon.

Patents have been recognised as a very rich data source for the study of innovation and technical change and there are many applications of patent analysis (see Lai et al., 2006). The innovative performance of organisations have been analysed with indicators such as research and development expenditures (R&D inputs), patents, patent citations and new product announcements. Increasingly researchers in technology management are using patent citations as an indicator of companies' innovation performance. One line of research counts the number of times a patent is cited in subsequent citations (forward citation) to measure its value or importance (Trajtenberg, 1990; Hall et al., 2005). A second line of research is interested in spillovers and knowledge flows and uses citations as indicators of knowledge transmission between inventors, firms and industrial sectors (Hall et al., 2001; Jaffe et al., 2000) The advantages and limitations (Table 1) of using patent data in economic research are widely discussed in the literature.

Patents are generally regarded as the precursors of technological development. Patent information is mainly intended for specialists, but especially their relations to other patents can make patents a valuable source of knowledge also for non-specialists trying to identify general trends. In this paper we first briefly define one special case of technological change, namely, the phenomenon of convergence. After introducing convergence, we provide evidence of the technological development in the printing and electronics industries. International patent classification (IPC) data is used to study whether there has been a growing overlap of the technological areas in which different industrial sectors are

Advantages and Opportunities	Limitations and Challenges

- Highly detailed information on the invention
- A homogenous measure of technological novelty and available for a long time series
- Includes citations of previous patents and the scientific literature
  - possibility to study spillovers
- evaluates the value of innovations
- evaluates the "originality" and "generality" of innovations
- A largely available stock of patents
- Data contained in patents are supplied in voluntary basis
- Possibility (regardless of the challenges) to integrate data into other complementary information (financial data, alliance data etc.)
	- Not all innovations are patented
	  - does not meet patentability criteria
	  - strategic decision to patent vs. other means of appropriability
	- Inter-industry and inter-firm differences in the propensity to patent
	- Filing patents under different names (e.g. subsidiaries)
	- Differences across countries in economic costs and benefits of patents
	- Interpreting findings of the citation analysis requires at least minimum knowledge of patenting search procedures and reports in different countries
	- "Truncation" problem in evaluating the importance of very recent inventions

Table 1. Advantages and limitations of patent analysis (Griliches, 1990; Hall et al., 2001; Michel & Bettels, 2001; Thoma & Torrisi, 2007)

operating, or whether technological profiles at the industry level remain distinct. Patent citation data is used to provide insights into the importance of technological transition of the paper and printing firms. Patent citations are used, because the value of patent counts as a proxy for R&D success is severely limited by the large variance in the significance of individual patents (e.g. Hall et al., 2005). All in all, the study has three main objectives. First, patents are used to find out converging technology areas between paper and electronics. Secondly, citations received are used to provide insights into future competition between the players. Thirdly, we analyse the pioneering innovations of the players in order to evaluate the breakthrough innovations of the players and discuss the opportunities and challenges of using patent data in the industry analysis.

The paper is structured as follows: Section 2 outlines the convergence phenomenon and characterises the emerging printed intelligence industry. Section 3 describes the data and methods used. Section 4 presents the empirical results and finally the study is concluded and discussed in Section 5.

2. Convergence evolution and the printed intelligence industry

2.1 Convergence

Convergent developments (Figure 1) and the emerging new industry segment between industries will potentially mean to fundamental changes leading to opportunities and challenges alike.

In the process of convergence, the technology bases of companies are becoming increasingly similar, which eventually means that companies compete with the same technological competencies. A new industry segment will either replace former segments or complement them at their intersection (Dowling et al., 1998; Bröring et al., 2006). In the "substitutive

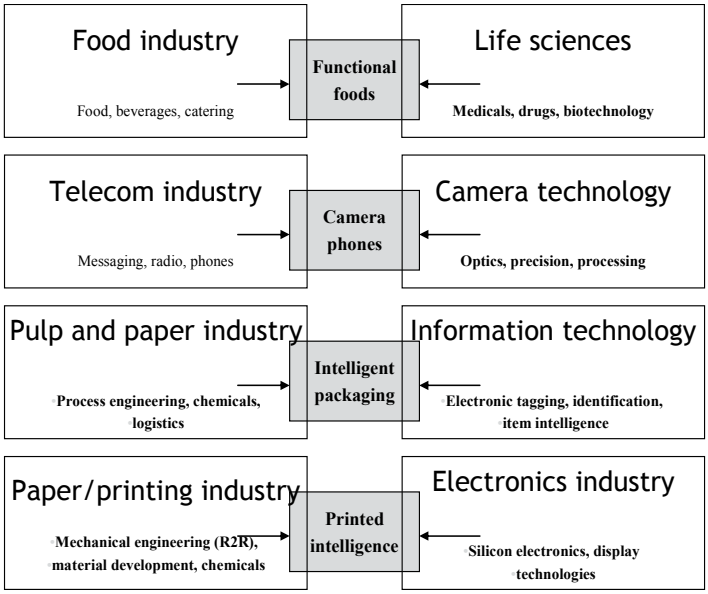


Fig. 1. Examples of convergent developments

paradigm" the new industry segment will replace the former segments (1+1=1) leading to competitive convergence. In the "cooperative paradigm" a new market emerges (1+1=3) that requires the combination of resources and competencies from previously separate industries (e.g. through strategic alliances or other forms of collaboration) leading to complementary convergence. (Dowling et al., 1998) In the "coopetitive paradigm" convergence may imply a need to collaborate and compete at the same time. All in all, convergence typically changes the basis for competitive advantage and firms must adapt their strategies depending on the nature and degree of convergence.

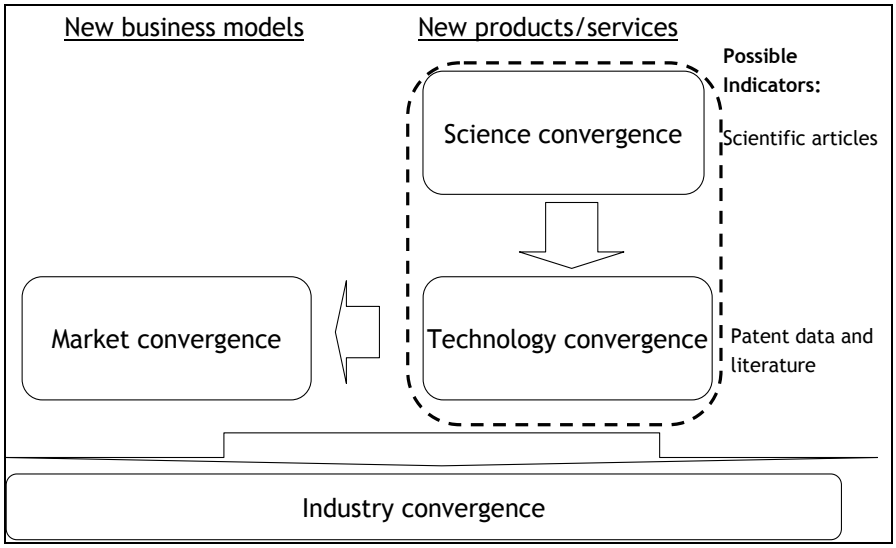


Fig. 2. Phases of convergence (Curran & Leker, 2010)

Curran & Leker (2009) base their concept of anticipating convergence upon the assumption of an idealised time series of events (Figure 2) starting with scientific disciplines, where distinct disciplines begin to cite each other and eventually develop further into closer research collaboration. After the distance between basic sciences has been decreasing, applied science and technology development should follow, leading to technology convergence. Then in market convergence, new product market combinations will emerge, and finally the fusion of firms and industry segments will lead to industry convergence.
Due to its high strategic importance, an early identification of trends of convergence and anticipation of changing industry structures matters to all stakeholders, including managers, academics and regulators. Anticipating and monitoring the stages of convergence process would enable firms to develop new competences, and form strategic alliances or acquire new technologies at the early stages. (Duysteers & Hagedoorn, 1998; Curran et al., 2010)

2.2 The RFID and printed intelligence industry

Printed functionality (or intelligence) means adding new functionalities into a flexible substrate, typically paper and plastics, in addition to regular graphical properties by using printing methods. Printed intelligence can be codes containing links to additional information. Such codes include one- and two-dimensional bar codes as well as invisible, reactive and electronic codes. Furthermore, printed functionality can be visual effects and images, electronics, optics and displays as well as sensors and indicators. The term *hybrid media* is related to printed functionality and is defined as the integration of different media, contents and functionalities, especially the convergence between fibre-based products and electronic media. Electronic paper belongs to the same context, even if the substrate here is plastics rather than fibre-based materials. Figure 3 shows how print media, electronic media, printed functionality and hybrid media interconnect. In the figure *printed* means that every part of a certain component is made by using printing methods, and *printed/attached* means that some parts of the component can be done by other means besides printing. One example is the radio frequency identification (RFID) tag where the antenna is printed and the chip attached to the printed antenna by other means. (Hakola et al., 2006)

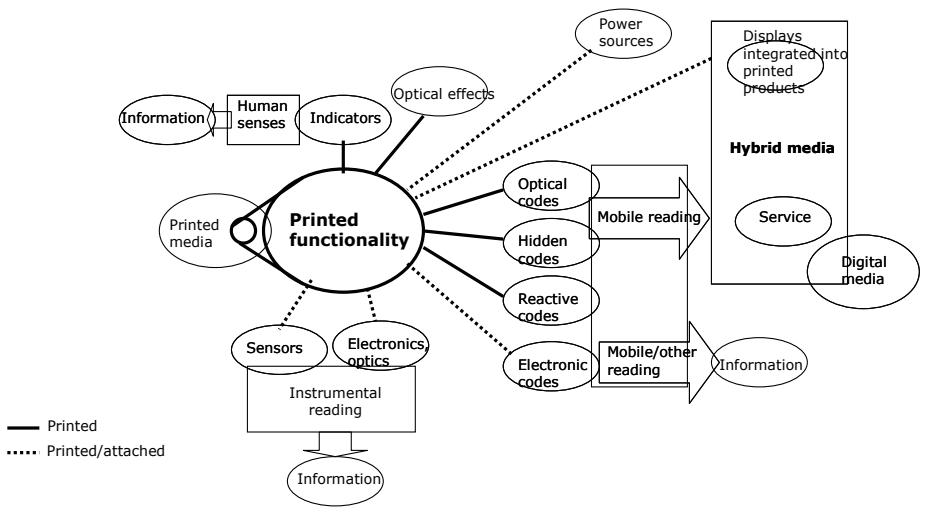


Fig. 3. The definition of printed intelligence and hybrid media (Hakola et al., 2006)

A smart or intelligent system can be defined to be one that has an inherent ability to gather information on its operating environment or history, to process information, to draw intelligent conclusions from it, and finally to act on those inferences by changing its characteristics in an advantageous manner. The most important development areas of intelligent packages can be defined to be concerned with product authenticity, anti theft or tamper evidence, track and trace, and brand enhancement. Possible technologies include electronic article surveillance (EAS), radio frequency identification (RFID), and other electronics or magic inks (IDTechEx, 2006). RFID technology can be regarded as the most mature smart development area (Aho, 2003).

RFID technology is a data collection and transfer technology that uses radio frequency waves to transfer data between a reader and an item. The RFID system consists of three basic components: a tag, a reader, and back office data-processing equipment. The tag contains unique identification information about the item to which it is attached; the reader emits and receives radio waves to read the information stored in the tag, and the data-processing equipment processes all the collected data (Wu et al., 2006). Tags can be active or passive: Active tags have a battery, whereas the passive tags use energy from the reader's signal to power-up the integrated circuit and transmit stored data back.

Today, most RFID tags contain a silicon chip and copper-etched circuit boards. This traditional technology sets limitations to production speed and capacity, and furthermore, the method is environmentally unfriendly. New roll-to-roll (R2R)-printing technology allows the use of a rotary-screen, lithographic, gravure or flexography press at a much higher production speed. Digital inkjet technology could also be used as an alternative method. At first, the antenna is printed and then the integrated circuit is attached or alternatively both are printed. The inlay is typically then placed on a label substrate. (Lynn 2005) Eventually, printed RFID (Lynn 2005; Harrop & Das 2008; Fortunato et al. 2008; Österbacka 2008) manufactured by efficient (R2R)- manufacturing method (Kesola 2007) is probably going to gain market share, and low-cost item-level RFID tagging could lead to peak in the industry sales.

According to Harrop & Das (2008), low cost flexible substrates are needed in order to open up new potential markets, since the applications of printed electronics are very price sensitive. The most popular substrates today are polyethylene thin films, but paper substrates offer low costs for processes, which can tolerate the rough surface of paper.

According to (Ngai et al. 2008), RFID systems have been applied, in particular, in supply chain management and manufacturing, but the potential application areas are much broader. Enterprises and entities today utilize RFID successfully in their every day operations for a wide variety of application areas. Most cases in the beginning of 2008 occurred in retail and consumer goods category, mainly because of mandates by major retailers and military organizations.

Passenger transport represented the second biggest application area. Many executed projects in the leisure application area came from one time events. Finance, security and safety sectors were also steadily growing and driven for example by counterfeiting and terrorism. Land and sea logistics is another steadily growing area. The total RFID market was worth about five billion US dollars in 2007, and it is forecasted to reach 27 billion US dollars in 2018, biggest potential being in East Asia, followed by North America and Europe. (Das & Harrop 2008) According to (Seppä & Uusikylä 2009) RFID could create a revolutionary innovation during 2010s in case Internet and mobile phones, capable to

communicate with RFID identifiers can be successfully interlinked. This could provide a possibility for myriad services literally to all products.
Electronic paper is a thin display technology designated to mimic the appearance of regular ink on paper. First commercial applications are used in electronic books as memory displays and advertisement signs in firms. Electronic ink refers to the field of technologies that can display persistent text and graphics and where the text and graphics are imprinted via use of electronic means. According to Aho (2003), conducting polymers have responses in their physical properties to various external stimuli, and another point of view in the combining conducting polymers and paper is printed electronics. Applications of conducting polymers include e.g. electronic components, optoelectronic devices and sensors. The aim of several research groups is to produce disposable, low-cost, and flexible laminar electronics produced using additive reel-to-reel techniques. With the help of printed electronics on paper the electronics industry could respond to the legislation demands of the future, and some laminar electronics have already been realised, such as the silicon transistor using reel-to-reel manufacturing, all-polymer transistor by screen printing, printed all-polymer field-effect transistor, all-organic printed rectifier, all-polymer RC filter circuit by ink-jet printing, and printed battery (Aho, 2003). Ink-jet printing seems to get more and more attention. Paper as a flexible substrate has been used in transistors, circuit boards and batteries with other chemicals than conducting polymers.

3. Methods and data

In order to identify a firm's technological domains, the observed IPC codes in the firm's patent records were identified and classified into technology fields representing the firm's major business domains. Patent application in each field indicates an accumulation of knowledge and advancement in the technological trajectory (Fai & Tunzelmann, 2001). IPC codes are a hierarchical way of assigning the category to which every patent belongs. There are 8 sections, 120 classes, 628 sub-classes and about 70,000 groups. In our analysis we have utilised a higher level classification, by which the 628 sub-classes are aggregated into 35 technological fields, and these are further aggregated into five main categories: Electrical engineering, Instruments, Chemistry, Mechanical engineering and Others. (Appendix 1, WIPO, 2008) From the paper & printing industry point of view we will call categories Chemistry and Mechanical engineering as "traditional fields" and Electrical engineering and Instruments as "emerging fields".

3.1 Citation analysis in convergent industry environments

The "backward citations" (i.e. citations made) in the patent document position the new invention technologically with respect to previous patents and "forward citations" represent citations received by the patent. Forward citations (citations made by other patents) are considered to reflect the patent's technological significance, the applicability and the ability of the inventors to benefit from their inventions, namely, their appropriability.
One aspect is to what extent patents are cited by the same assignee (we refer to these as self-citations) as presumably citations that belong to the same assignee represent transfers of knowledge that are mostly internalised. Self-citations would suggest that the firm has a strong competitive position in the particular technology and is in a position to internalise the knowledge created by its own developments. Hall et al. (2005) found that patent citations could provide a more accurate picture of the company's intangible assets and in

particular, the value potential captured seemed to be enhanced in settings where forward citations are made by the inventor.

We distinguish self citations from external citations, and further divide external citations into two groups: within the industry and beyond the industry citations (Figure 4).

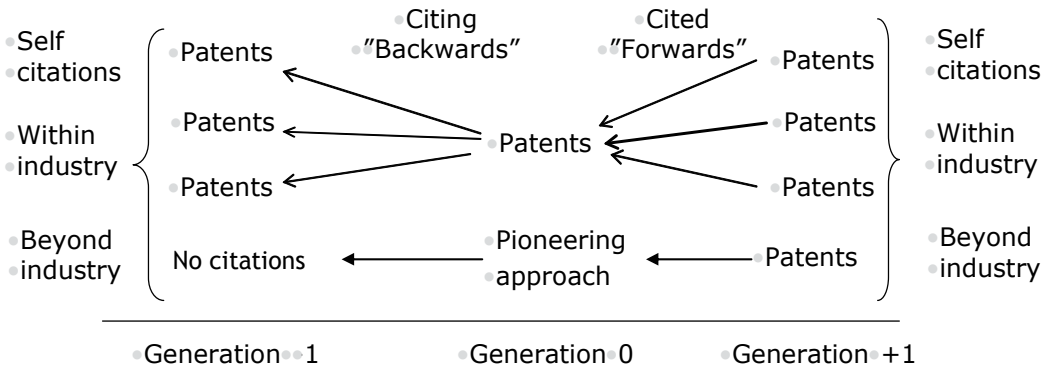


Fig. 4. Different types of citations

Self citations typically imply a more competitive position in that field (capability), external citations may suggests that the citing firm is entering a technological competition within (competition) or beyond (diversification) industry. Generally we expect that in convergent environments spill-over effects from extra industry increases and the gradual diversification to the new fields is evolving. Forward citations provide indications about the future success of these new inventions and the future competition between the players. In addition we can distinguish the inventions which have not made any backward citations, but have received many citations, which we call "pioneering innovations".

3.2 Sample and databases

Our analysis is based on the links between two datasets. The primary source for acquiring the list of companies and their activities in the RFID field was an independent consultancy company (Das & Harrop, 2008), and the companies were in further analyses categorised into four groups: value chain's upstream focused players, downstream focused players, vertically integrated firms, and paper and printing companies. Upstream players are involved in developing, manufacturing and selling identifiers such as chips, antennas, tags and labels, or devices such as readers and printers. Downstream focused firms are involved in software and integration, and they operate closely in the end customer interface. Vertically integrated firms operate broadly in the whole value chain with activities both upstream and downstream of the value chain. In addition we separately analysed printing and paper companies operating in the field. The concordance, based on Smoch et al. (2003), between paper and publishing and electronics industry classification and the companies evaluated are presented in Appendix 2.

Information on patenting was drawn from the EPO Worldwide Patent Statistical Database. The coverage of the database regards documents from more than 80 patent offices worldwide since the 1970s. This worldwide statistical patent database, also known as PATSTAT, was developed by the EPO in 2005 and first published in April 2006. The elements of PATSTAT are the title and abstract of application, filing, priority and

publications dates of the application, applicants and inventors and detailed addresses, the IPC classification symbol and priority applications. Moreover, PATSTAT provides complementary information, for example, on the citation links such as the category of the citation, citation identification, origin of the citation and non-patent literature bibliography. (EPO, 2008) In this study we used these data mainly to identify (1) in which IPC classes the players have patented and to find converging technology areas, (2) forward citations of patents in order to evaluate the impact and value of innovations in the converging industry sectors and (4) the pioneering innovations of the players.

4. Empirical results

4.1 Patenting in traditional and emerging fields

In the analysis there were altogether 87 firms which were categorised into four different clusters under the following headings: upstream focused players (N=26), vertically integrated players (N=23), downstream players (N=17) and paper & printing companies. In the empirical part we analysed each cluster patents and their forward citations in the years 1960–2006. The analysed firms had altogether 464,225 patent applications and the top 50 IPC classes in each cluster. The players' (Table 2) patent distribution clearly shows that downstream players have the most focused technological competencies as 95.8% patents are related to electrical engineering patents and the paper and printing firms have the most diversified patent portfolio.

Industry / IPC group	Paper & Printing (N=18)	Upstream electronics (N=26)	Vertically integrated electronics (N=23)	Downstream electronics (N=17)
Patent count	77,963	124,184	218,560	43,518
TOP50 IPC (%/all)	84.7%	87.6%	91.7%	98.0%
I Electrical engineering	22.3%	80.8%	88.7%	95.8%
II Instruments	16.5%	16.2%	6.9%	3.3%
III Chemistry	18.5%	0.5%	1.6%	0.0%
IV Mechanical engineering	41.4%	2.5%	2.9%	0.9%

Table 2. The players' patents distribution to the technological fields 1978–2006

Paper and printing firms have traditionally been strong in mechanical engineering and chemistry patents. In the sample, 41.4% of the paper and printing firms' patents were related to mechanical engineering. Figure 5 suggests that there has been a change in the focus for the paper and printing industry in the recent years. With the emergence of printed intelligence and RFID there has been a significant increase in patenting in the new fields.

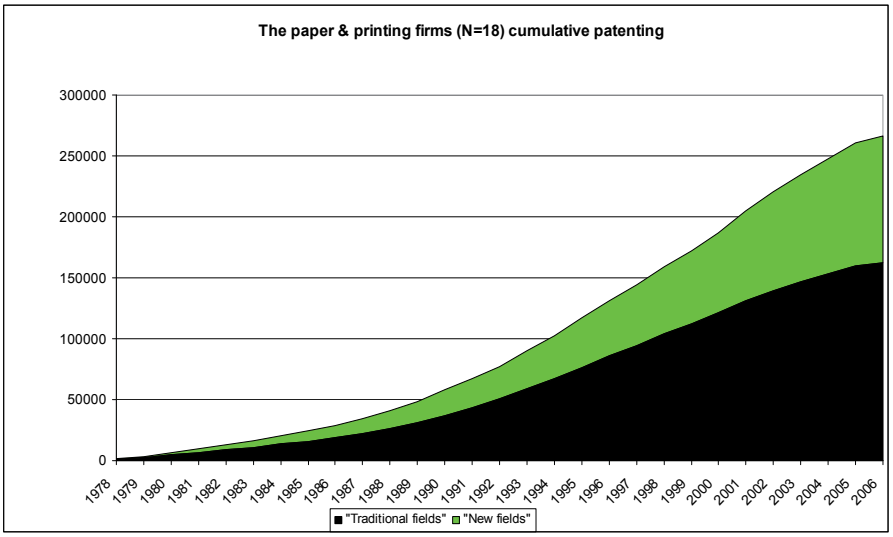


Fig. 5. The paper & printing firms cumulative patenting in traditional and emerging fields

4.2 Citations received and dominance of emerging technologies

The analysed firms had received altogether 1,389,224 citations to the patents, of which 144,479 (10.4%) were self-citations. Table 3 shows that vertically integrated firms have received overwhelmingly most of the citations with an average of over four citations per patent. The paper and printing firms have received only 0.91 citations per patent. We analysed TOP50 IPC classes citations received for each cluster and their distribution to the technological fields.

Industry / IPC group	Paper & Printing (N=18)	Upstream electronics (N=26)	Vertically integrated electronics (N=23)	Downstream electronics (N=17)
Patent count (Citations received)	70,847	292,595	904,335	121,447
Citations received (average)	0.91	2.36	4.14	2.79
TOP50 IPC (%/all)	76.4%	86.0%	89.7%	96.3%
Self-citation count	5,951	29,260	96,764	13,237
(%)	(8.4%)	(10.0%)	(10.7%)	(10.9%)
- Within industry	71.4%	86.0%	89.3%	95.8%
- Beyond industry	29.6%	14.0%	10.7%	4.2%
External citations	64,896	263,335	807,571	108,210
	(91.6%)	(90.0%)	(89.3%)	(89.1%)
- Within industry	59.0%	79.9%	84.6%	91.9%
- Beyond industry	41.0%	20.1%	15.4%	8.1%

Table 3. The players' forward citations distribution to the technological fields 1978–2006

The paper and printing firms' forward citations distribution between the technological fields shows a declining trend in the traditional fields, whereas the significance of electrical engineering patents has steadily increased. Many of the electrical engineering citations received is related to audio-visual technology, computer technology and semiconductors. The biggest growth in citations received has been in semiconductor device patents, indicating a growing competition especially with upstream electronics players. Regardless of the fact that as a whole vertically integrated players seem to be technology leaders, the paper and printing firms beyond industry self-citations indicate quite a strong capability development in new fields and received external citations indicate also market some market power in new fields.

The dominance of emerging technologies

The patent analysis reveals that from the paper and printing industry point of view the most important emerging technology fields are related to the computer technology (G06K; G06F), audio-visual technology (G09F; G11B; H04N; H05K); semiconductors (H01L), and optics (G02F; G02B; G03F, G03G[1]. Vertically integrated players have made most of the patents in computer technology, audio-visual technology and semiconductors, and the paper and printing firms have made most of patents related to optics. In the computer technology the downstream players have significantly increased their share (Figure 6) and recently even surpassed the integrated players. Integrated players have made very much self-citations in the computer technology indicating a very strong capability position in these technologies. The downstream players have also lately made a lot of self-citations in these technologies indicating an intensifying competition between the players. The paper and printing companies has quite a marginal share of computer technology patents.

Integrated players dominate also audio-visual technology patents, but the downstream have taken competitive position also in these technologies in the 2000's. Upstream players have taken stable 15% share of patents throughout the period, while in the paper and printing companies share have been moderately increasing. Interestingly, integrated players have made most of the self-citations and also upstream have made much self-citation, while there have been only marginal increase in the downstream players' self-citations.

In the semiconductor patents integrated players are strong, but upstream players are taking the dominance in the 2000's. In the optics patents (Figure 7) the paper and printing firms have made most of the patents and increased relative share throughout the period.

Table 4 shows a data of the patents and citations received to the emerging fields. The high figure of citations received of integrated players indicates that the patents have been technologically and economically significant. On the contrary the low figures of citations made and received by the paper and printing firms indicates that the spillover effect from emerging fields have not been so tremendous and the importance of new inventions have not been, at least so far, so extensive compared to other players. Of course, when

[1] G06K - Recognition of data; presentation of data; Record carriers; Handling; G06F Electric digital data processing; G11B - Information storage; H04N - Pictorial communication; H05K - Printed circuits; H01L - Semiconductor devices; G02F - Devices or arrangements; G02B - Optical elements, systems, or apparatus; G03F - Photomechanical production of textured or patterned surfaces; G03G – Electrography; electrophotography; magnetography

interpreting these results we have to remember the inter-industry differences in patenting practices. Downstream players have very focused technological competencies related to computer technology, and the citations indicate that the few patents in other fields seem to have been significant.

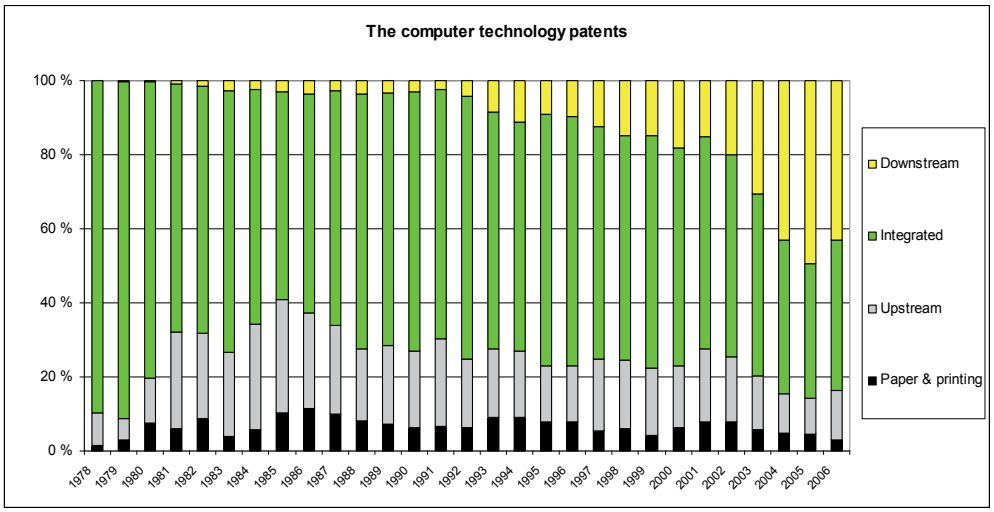


Fig. 6. The relative share of the computer technology patents 1978–2006

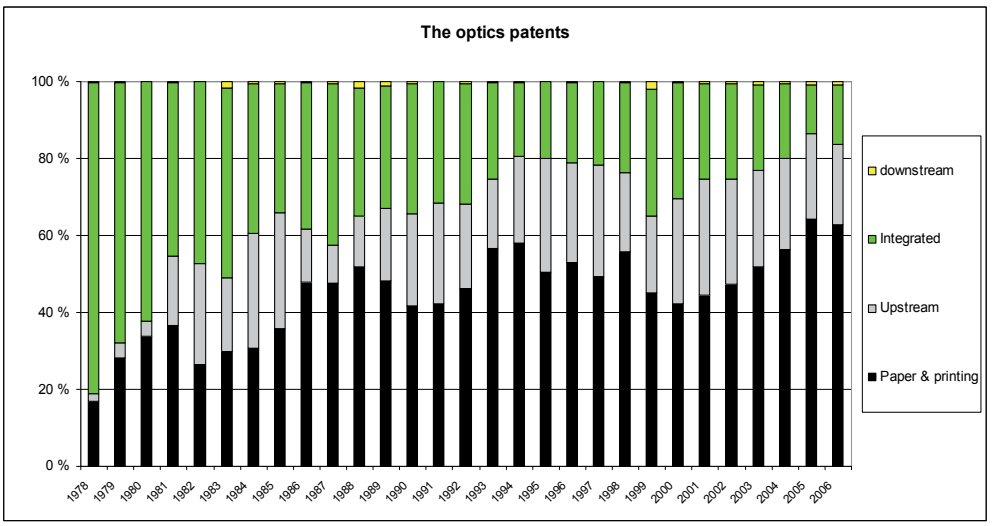


Fig. 7. The relative share of the optics patents 1978–2006

Industry / IPC group		*Paper & Printing (N=18)*	*Upstream electronics (N=27)*	*Vertically integrated electronics (N=24)*	*Downstream electronics (N=18)*
Patents (IPC4) 1978–2006	*Computer technology*	8,970	23,701	80,709	29,516
	Audio-visual technology	6,685	10,410	32,315	3,650
	Semiconductors	5,224	29,519	29,761	241
	Optics	16,863	8,036	9,926	204
Citations received/ Patent Self-citations (%)	*Computer technology*	0.67 6,007 (5.2%)	2.80 66,343 9.9 %	4.51 364,230 (14.8%)	2.78 82,245 (12.8%)
	Audio-visual technology	0.95 6,319 (8.7%)	2.47 25,745 (7.7%)	4.47 144,502 (9.2%)	3.25 11,847 (5.1%)
	Semiconductors	0.77 4,000 (4.0%)	2.91 86,032 (13.3%)	4.92 146,562 (11.9%)	6.50 1,567 (1.8%)
	Optics	0.79 **13,384** (10.3%)	3.23 **25,919** (7.2%)	3.98 **39,456** (5.9%)	5.02 1,024 (6.8%)

Table 4. The players' patents and forward citations in converging technology areas

In the semiconductor patents the integrated and upstream are the strongest players. Relatively the upstream players have increased their share mostly and have taken the dominant position and increased for example their self-citations significantly. However, in the citations received the integrated players have stayed as the dominant player. Paper and printing companies have taken the dominant position in patents related to optics. In the citations made, however, integrated and upstream players have made substantially more citation than the paper and printing companies. Integrated dominates also the citations received, whereas upstream and paper and printing companies have increased their share of forward citations. Paper and printing companies' self-citations to optics technologies, however, have increased throughout the period, indicating a strong competitive position in these fields.

4.3 The pioneering innovations

The distribution of pioneering innovations (Table 5) is quite similar compared to patent distribution, and so the players' have these more radical innovations mostly to their own strong technological fields.

Vertically integrated players have made most of the pioneering innovations, followed by upstream and paper and printing industry players. The paper and printing firms pioneering innovations have increased especially in optics, semiconductors, computer technology and in the basic communication processes related technologies, but compared to the other players paper & printing companies had make substantially less pioneering innovations than the other industry players. In the sample there were totally 306 inventions which have received at least 100 citations, and from these innovations over 60% were made by integrated players. When looking ten most cited pioneering innovations (Table 6) we can see how dominant IBM have been in these breakthrough innovations.

Industry / IPC group	Paper & Printing (N=18)	Upstream electronics (N=26)	Vertically integrated electronics (N=23)	Downstream electronics (N=17)
All patents	77,963	124,184	218,560	43,518
Pioneering innovations (self-citations)	16,091 (18.9%)	26,855 (31.7%))	72,461 (46.1%))	8,459 (50.9%)
Citations received / patent	2.72 (43,747)	5.36 (143,829)	3.42 (247,834)	9.40 (79,546)
TOP50 IPC (%/all)	83.0%	83.7%	90.1%	98.1%
I Electrical engineering	**26.1%**	**75.0%**	**88.8%**	**95.9%**
II Instruments	**19.8%**	**21.7%**	**8.1%**	**3.6%**
III Chemistry	**17.3%**	**0.6%**	**1.0%**	**0.0%**
IV Mechanical engineering	**36.8%**	**2.7%**	**2.1%**	**0.5%**

Table 5. The players' pioneering innovations distribution

Position	Applicant	Year	Technology field	Count of cites
1	1 IBM	1988	Semiconductors	434
2	IBM	1996	Computer tech.; semiconductors	412
3	IBM	1985	Semiconductors; audio-visual technology	390
4	Texas Instruments	1992	Telecommunications	376
5	IBM	1993	Machine tools; measurement; semiconductors	344
6	IBM	1990	Measurement, audio-visual technology	338
7	IBM	1990	Computer technology; Digital communication	328
8	IBM	1994	Computer technology; control; semiconductors; telecommunications	320
9	IBM	1996	Computer technology	315
10	IBM	1989	Computer technology	304

Table 6. The ten most cited pioneering innovations

From the upstream players Texas Instruments and Symbol Technologies have made most of their pioneering innovations. Microsoft has clearly dominated the downstream players pioneering innovations, and Moore, Toppan printing, Weyerhaeuser, Dai Nippon and Avery Dennison have been strong paper & printing companies in pioneering innovations. Interestingly, when looking also cites of second generations citations (Table 7), there seem to be huge variation between the first and second generation citations.

Applicant	Year	Technology field	Count of cites	Cites 2nd generation
Integrated/IBM	1988	Semiconductors	434	2671
Integrated/IBM	1996	Computer tech.; semiconductors	412	2442
Integrated/IBM	1985	Semiconductors; audio-visual technology	390	5672
Upstream/Texas Instruments	1992	Telecommunications	376	5268
Upstream/Texas Instruments	1994	Computer tech.; control; digital communication	243	1230
Upstream/Texas Instruments	1984	Micro-structural and nano-technology; optics	239	2732
Downstream/ Microsoft	1996	IT methods for management; digital communication	276	2625
Downstream/ Tibco Inc.	1990	Computer tech.; digital communication	270	2931
Downstream/ Microsoft	1996	IT methods for management	245	2081
Paper & printing Moore	1991	Computer technology; audio-visual technology, digital communication	127	2961
Paper & printing Moore	1996	IT methods for management	114	600
Paper & printing Toppan printing	1993	Semiconductors; electrical machinery, apparatus, energy	113	703

Table 7. TOP3 cited pioneering innovations of the each player

The patents which have been cited both in first and second generation can be thought to be a real breakthrough and long lasting innovations, while some of the pioneering patents seem to be superseded quite quickly with some new innovations.

5. Disscussion and conclusions

The global intellectual protection system is a rich source of valuable technologies as well as signals of technical emergence. Citations of the common patents may indicate the

convergence of technological competencies between firms in different industries toward solving a common problem or exploiting a common technology. The following of particular industrial sector patents may be detected as citations that allow both backward and forward searching from the patents. Such searching not only reveals inventors and assignees that may be valuable partners in technology assessment, but also cross-disciplinary citation and the linkage of fields through co-citation of a patent or technological field. Cross-fertilisation of ideas from other technical fields is frequently a rich source of new innovations in both basic science and the commercial sector (See Winter, 2000). In the paper & printing and electronics industries, important technological innovations are moving them closer together as more cross-scientific research can enable the printing sector to utilise technological developments in the neighbouring disciplines. Patents and forward citations were used in trying to evaluate the significance of this industry transformation. The downside of using forward citations in evaluating the technological significance and the economic value is that they are not available until a substantial period after the granting of a patent, because time is needed to accumulate significant information about its citations. In practice this means that the analysis will be challenging for the evaluation of current or very recent innovations. In comparison, backward citations provide comparable information upon publication of the patent document and, consequently, they provide comprehensive results earlier. Differentiating between external and self-citations within and beyond industry citations aids to provide more comprehensive prospects of future industry evolution. In addition, self-citations typically indicate a strong competitive position in the particular technology and the firms are in a position to internalise the knowledge created by their own development. The patents which have not been made any citations of previous patents (no prior art), but have been cited a lot (forwards) are called as pioneering innovations. These four different kinds of citations are used as a tool to find out the connection between technology development and trajectory changes in convergent environments.

The patent data were collected from 87 main players operating in the RFID value chain. In the empirical analysis, over 465,000 patents and their over 1.3 million forward citations were analysed in the years 1978–2006. The results of the study indicate that paper and printing companies still patent predominantly in mechanical technologies and chemicals, but increasingly in electronics technologies, suggesting that the industries are becoming more technologically convergent. The growing overlap of the technological fields in which different industrial sectors are operating show clear indications for convergence and strategically important knowledge of the future competitive area between are presented. This type of patent analysis helps companies to recognise trends early in the industry and take strategic decisions accordingly.

It seems evident that the importance of external innovation will increase and the winners will be those who succeed to capture external innovations from outside the company and learn to use collaborative R&D. So it seems that the essential knowledge from beyond one's own industry is necessary and key to innovation management. Considering the whole electronics industry, the paper & printing companies are still quite marginal players, although potentially quite huge new markets should be available when the technology progresses to low cost flexible substrates. Moreover, the self-citation analysis indicates that the paper and printing firms have received self-citations mainly in their traditional fields of core competencies and there have been not so much pioneering innovations compared to the players. Self-citations would then suggest that the firms have not yet a strong enough

competitive position in the new fields. All in all, we see that patent citation analysis and convergence should be included in the research agenda of technology management and firm strategy.

6. Acknowledgements

The paper is an extended version of the 6th International Working Seminar on Production Economics, Innsbruck, March 1-5, 2010. The Authors would like to thanks conference participants for their valuable feedback and comments. In addition we want to thank Juha Kortelainen for his support in data analysis of this paper.

7. References

Aho, O. (2003), Combining conducting polymers with paper surface or fibres, Helsinki University of Technology, Laboratory of Paper Technology, Report, Functionalized paper (PAPU) project.

Bröring, S., Cloutier, M. L. & Leker, J. (2006). The front end of innovation process in an era of industry convergence: evidence from nutraceuticals and functional foods, *R&D Management*, Vol. 36, Issue 5, pp. 487-498, 2006.

Curran, C., Bröring, S. & Leker, J. (2010). Anticipating converging industries using publicly available data, Technological Forecasting & Social Change, Vol. 77, Iss. 3, pp. 385-395.

Curran, C., Leker, J. (2009). Seeing the Next iPhone Coming Your Way: How to Anticipate Converging Industries, PICMET 2009 Proceedings, August 2-6, Portland, Oregon USA.

Das, R., Harrop, P. (2008). RFID forecasts, players & opportunities 2008-2018, IDTechEx Ltd. Cambridge.

Dowling, M., Lechner, C., Thielmann, B. (1998). Convergence: innovation and change of market structures between television and online services, *Electronics Markets Journal*, Vol 8. No. 4. pp. 31–35.

Duysteers, G., Hagedoorn, J. (1998). Technological Convergence in the IT Industry: The Role of strategic technology alliances and technological competencies, *International Journal Economics of Business,* 5(3), 355-368.

EPO Worldwide Patent Statistical Database, April 2008.

Fai, F., Tunzelmann von N. (2001). Industry-spesific competencies and converging technological systems: evidence from patents, *Structural change and Economic Dynamics*, Vol. 12, pp. 141-170, 2001

Fortunato, E., Correia, N., Barquinha, P., Pereira, L., Goncalves, G. & Martins, R. (2008). High- Performance Flexible Hybrid Field-Effect Transistors Based on Cellulose Fiber Paper. *IEEE Electron Device Letters*. Vol. 29 Issue 2, pp. 988-990.

Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature*, Vol. 28, pp. 1661-1707.

Hacklin, F. (2008). Management of Convergence in Innovation – Strategies and Capabilities for Value Creation Beyond Blurring Industry Boundaries. Physica-Verlag, Heidelberg, 2008.

Hakola, L., Lindqvist, U., Linna, H., Siivonen, T. & Södergård, C. (2006). Roadmap on printed functionality and hybrid media. 33rd International Research Conference of IARIGAI. Sept. 10-13, Leipzig, Germany.

Hall, B. H., Jaffe, A. B., Trajtenberg, M. (2005). Market value and patent citations, *RAND Journal of Economics*, Vol. 36 (1), 16-38.

Hall, B. H., Jaffe, A. B., Trajtenberg, M. (2001). The NBER Patent Citations Data File: Lessons, Insights, and Methodological tools, NBER Working Paper No. 8498, Cambridge, MA.

Harrop, P. & Das, R. (2008). Introduction to Printed Electronics. IDTechEx Ltd. Cambridge, United Kingdom.

Jaffe, A. B., Trajtenberg, M., Fogarty, M. S. (2000). The Meaning of Patent Citations: Report on the NBER/CASE- Western Reserve Survey on Patentees. NBER Working Paper No. 7631, Cambridge, MA.

Kesola, I. (2007). Roll to Roll, cost effective integration of high-tech technology and traditional production methods. Pulpaper 2007 Conference, June, Helsinki. Finland.

Lai, K., Lin, M., Chang, S. (2006). Research Trends on Patent Analysis: An Analysis of the Research Published in Library's Electronic Databases, The Journal of American Academy of Business, Cambridge, Vol. 8, No. 2., pp. 248-253.

Lei, D. T. (2000). Industry evolution and competence development: the imperatives of technological convergence, *International Journal of Technology Management,* Vol. 19, pp. 699-738.

Lind, J. (2004). Convergence: History of term usage and lessons for firm strategists, In: ITS 15th Biennial Conference, Berlin, Germany. International Telecommunications Society (ITS), 2004.

Lynn, C. (2005). RFID and Printed Electronics: A new Opportunity for Printers?. Analyzing Publishing Technologies. Seybold Report: Analyzing Publishing Technologies. Vol. 4 Issue 24, pp 14-17.

Michel, J. & Bettels, B. (2001). Patent citation analysis: A closer look at the basic input data from patent search reports, *Scientometrics*, Vol. 51, No 1, pp. 185-201.

Ngai, E., W., T., Moon, K., K., L., Riggins, F., J. & Yi, C., Y. (2008). RFID research: An academic literature review (1995-2005) and future research directions. *Journal of Production Economics.* Vol 112 Issue 2, pp 510-520.

Pavel, P., Pavitt, K. (1997). The technological competencies of the world's largest firms: complex and path-dependent, but not much variety. *Research Policy,* Vol. 26, pp. 141-156.

Seppä, H. & Uusikylä, M. (2009). Vallankumouksellinen RFID. Etätunnistusteknologian kehitys meillä ja maailmalla. Tekesin katsaus 249/2009. Tekes. Helsinki. Finland.

Smoch, U., Laville, F., Patel, P., Frietsch, R. (2003). Linking Technology Areas to Industrial Sectors, Final Report to European Commission, DG Research.

Stieglitz, N. (2003). Digital Dynamics and Types of Industry Convergence: The Evolution of the Handheld Computer Market, In Christensen, Jens F. (eds.) The Industrial Dynamics of the New Digital Economy. pp. 179-208.

Suoranta (2008). Open Modular Platform Architecture – A Key Enabler for Open Innovation. XIX ISPIM (International Society for Professional Innovation Management) Conference, Tours, France.

Thoma, G., Torrisi, S. (2007). Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases. CESPRI Working Papers 211.

Trajtenberg, M. (1990). A Penny for Your Quotes: Patent Citations and the Value of Innovations. *The Rand Journal of Economics,* Vol. 21 (1), 172-187.

Winter, S. (2000). Appropriating the Gains from Innovation. In. Day, G. S., Schoemaker, P. J. H., & Gunther, R. E. (eds.), Wharton on Managing Emerging Technologies. John Wiley & Sons, New York, NY.

Wirtz, B.W. (2001). Reconfiguration of Value Chains in Converging Media and Communications Markets, *Long Range Planning,* Vol. 34, Vol 4, pp. 489-506.

World Intellectual Property Organization (2008). World patent report: A statistical review.

Wu, N.C., Nystrom., Lin, T. R., Yu H. C. (2006). Challenges to global RFID adoption. *Technovation,* Vol. 26, no. 12, pp. 1317-1323.

Österbacka, R. (2008). Intelligence in Printing. Signs of Renewal in the Forest Industry. Summer School 2008, Lappeenranta University of Technology, 9-10 September. Lappeenranta

5

Simulation Modelling of Manufacturing Business Systems

Nenad Perši
University of Zagreb, Faculty of informatics and organization Varaždin
Croatia

1. Introduction

Due to their intense dynamics, business environment relationships in the market drive constant changes within business systems and continuous accommodation to the newly arisen circumstances. As business operations have become too complex and dynamic, adaptability has arisen as one of the major features of business systems. A possible solution to the problem of increased flexibility and the ability for quick adaptation is the development of a business system states model as well as an alternative states management system model. It must be noted that the state and alternative solutions management system needs to respond timely and appropriately, taking into consideration the integrity of business processes and available resources.
Contemporary information and communication technologies (ICT) have established themselves as a means of providing business systems support primarily due to their technological capability of processing vast amounts of information in a short period of time. Such processing is aided by scientific methods and techniques of business system modelling and design. Consequently, ICT and the solutions generated by them have become essential forms of business systems management support as well as the basis for the improvement of the efficiency of business systems themselves.

2. Features of complex business manufacturing systems

The success of a business system in a competitive market depends on, among other factors, the internal organization of the system. It is for that reason that business activities performed within the business system are integrated into units called business functions. Generally speaking, each business manufacturing system consists of three basic business functions: finance, marketing and production. Whereas the production function can be treated as an individual function, marketing and finance functions are considered as part of logistic support to production. The processes and activities that are conducted within the finance business function are related to the management of a business system's financial resources. From the perspective of a business system's input this is reflected in the procurement of equipment, raw materials and other materials; in the course of production financial resources are provided for the operational costs incurred; finally, when output is concerned, they are directed at investment returns that include the profit enabling further growth and development of the business system. The role of the financial function is to

ensure a balanced and unhindered execution of processes and activities that occur within a business system regardless of the dynamics of the inflows and outflows of financial resources. The purpose of that function is fundamentally the same in unprofitable business systems too, although the manner in which resources are managed is somewhat different from that in profitable systems.
Marketing is the business function that ensures the development and improvement of products manufactured and services provided by a business system. Marketing and sales activities have an important role in finding customers and identifying their needs as well as in educating buyers and convincing them to use particular goods and services. The procurement part within the marketing business function ensures the purchasing of the required raw materials and other resources. Along with the finance function, its role is to provide conditions for initiating business processes and maintaining their constant development and growth.
The production function (production) is a set of business processes that create a particular material product or a service. Determinants of production are unique to each business system, with specific inputs, processing and outputs. Nevertheless, it is always connected with other business functions. This function is responsible for transforming inputs into certain outputs recognizable as a specific finished product. Moreover, this transformation is aimed at creating added value through finished products which can subsequently meet customers' needs. Without the added value, such a business system could be excluded from the supplier-manufacturer-buyer chain, as the buyers would be able to satisfy their needs directly from the supplier. Therefore, it is evident that creating added value is the essence of production. In business systems whose business processes are not based on processing materials (e.g. schools) creating an added value is measured by means of indicators which refer to socially useful values.

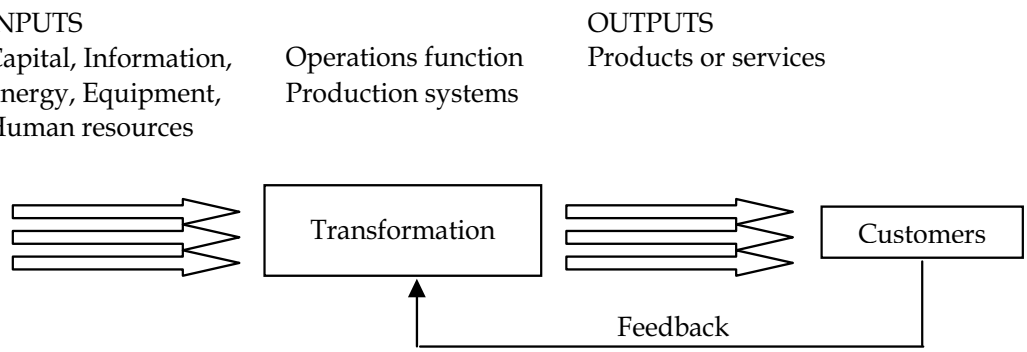


Fig. 1. Conceptual model of a business manufacturing system (Dilworth,1999)

2.1 Types of business manufacturing systems

Generally speaking, there is a large number of business manufacturing systems, each of them based on its specific product. However, there are certain features that they have in common. At the highest level of generalization, business systems can be divided into service and commodities manufacturing business systems (Dilworth,1999). However it is not easy to draw a clear line even at that level as contemporary business systems comprise both types of business activity (e.g. restaurants). The division into service and commodities

manufacturing business systems is based on the fact that in case of service business systems the user is involved in the business system itself, whereas in commodities manufacturing systems business processes occur in separate processing locations that the user commonly does not participate in (Dilworth,1999). There are several general differences between service and commodities manufacturing business systems (Dilworth,1999):

- Productivity is easier to measure in commodities manufacturing business systems since in manufacturing 'tangible' goods are processed, while the evaluation of services is fairly difficult.
- Quality assurance is easier in commodities manufacturing business systems: in case of commodities it is possible to directly verify whether certain quality norms are observed in the manufacturing process, while service quality evaluation may be more subjective and depends on an individual user.
- In service business systems there is a direct contact between the personnel and the user, while in commodities manufacturing systems contacts between the personnel that manufacture the product and the user are not common. Relations with customers are more important in service business systems since the quality and type of service directly depend on that relationship. The marketing business function of commodities manufacturing business systems also includes customer relations, although it does not directly affect the execution of the manufacturing system.
- Commodities manufacturing business systems can use warehouses as support to their activity. On the other hand, for service business systems warehousing is not available so their productivity depends exclusively on the disposable time and work resources (service duration time, number of workplaces etc.). As a result, service systems use time sharing to minimize the number of rejected requests.

2.1.1 Service business systems

In service business systems the customer participates in business processes. There are several subtypes of service business systems which have certain characteristics in common. The first of them is the level of participation of customers in a business process. It refers to the character of the mutual relationship between the customer and the service system, that is, to the amount of time dedicated by the system to each particular customer. Consequently, this type of service can be viewed from the perspective of two levels of participation – individual and collective. The second feature is the level of complexity, according to which service systems are divided by the type of specific knowledge and skills, type of equipment or the quantity of material and financial resources needed for the execution of a job. As a result, services classified by this feature can manifest a higher or a lower level of complexity. Both features are mutually related so certain services can be classified according to both categories.

2.1.2 Commodities manufacturing business systems

Depending on the nature of the system, commodities manufacturing business systems can be divided into the following categories, according to the character of the manufacturing process:

- job shop
- repetitive manufacturing
- batch manufacturing.

Job shop is the term which refers to manufacturing of unique custom-made (or even hand-made) articles. Its distinctive features are a small quantity of manufactured items within a wide range of products aimed at increasing sales probability and a fairly high price. This calls for the procurement of general purpose equipment and employees with a broad knowledge base possessing a variety of skills. Manufacturing itself must be extremely flexible, unrestricted by a firm business structure and a continuous flow process. Since manufacturing planning, distribution and coordination are determined by the momentary situation, extraordinarily skilled management staff is generally implied in this manufacturing type.
Repetitive manufacturing is the term which refers to manufacturing of large quantities of identical or similar articles. This manufacturing type is distinguished by a firmly defined organizational structure with a predetermined business flow. Business processes are sets of linearly connected short-term activities, each of them different in relation to all the other activities involved. The outcome of such organization is workplaces equipped with special tools and machinery requiring highly-specialised staff. Material management and raw materials management presupposes an input and an output warehouse as the sole prerequisite for a continuous production flow. Provided all the conditions are fulfilled, the process of manufacturing a single item is very short indeed.
Batch manufacturing seems to be a compromise between the job shop and repetitive manufacturing. It is the prevailing manufacturing type in manufacturing systems. Batch manufacturing is distinguished by defining initial requests for manufacturing a particular product, with the process being finalized after it has been repeated a specified number of times. For each product manufacturing equipment is adjusted and process flows and manufacturing structure are redefined. Such an organizational structure of the system and equipment should be flexible enough to be easily adapted for new jobs. The key parameter in those changes is the time needed for the manufacturing process adjustment. The scope of the required changes, which implies a shorter time needed for adjustments, can be reduced by focusing the manufacturing to similar types of goods. This can result in lesser changes in organizational and material flow, the need to use specialised equipment and a possibility to automate production.

3. A glass container manufacturing business system

Manufacturing business systems, as a particular business systems class, share characteristics related to business processes. This implies that certain business processes can be unified and shown by means of a general business processes model for manufacturing business systems. By recognizing business processes and their individual characteristics it is possible to determine and describe system states, conditions and mechanisms of system states changes.
Manufacturing glass containers is an example of a commodities manufacturing business system. Due to its characteristics, it can be classified as a batch manufacturing system. This concrete manufacturing system is used in this paper as the basis for the development of a model of an analogous business systems class.
A glass container manufacturing system belongs to a group of systems combining continuous and discontinuous manufacturing (Bider, 2005). The product line of the manufacturing system observed is hollow container glass varying in size, shape, colour and purpose. Within a glass container manufacturing system two different types of manufacturing processes are combined. The first part, called the hot zone, is where glass

mass to be later used for glass container production is prepared and melted. This manufacturing phase represents the process-type (closed-type) continuous manufacturing of standard products. Other manufacturing operations, ranging from the molten glass separation to all the cold zone operations, are performed discontinuously using different types of machinery and equipment on each particular item within the product line. The product type is determined by the glass mix itself and the kind of glass-forming tool used, so this segment of production, owing to its features, represents the chain-type (open-type) continuous manufacturing of standard products. A basic outline of the functioning of the observed business manufacturing system is shown in Fig 2.

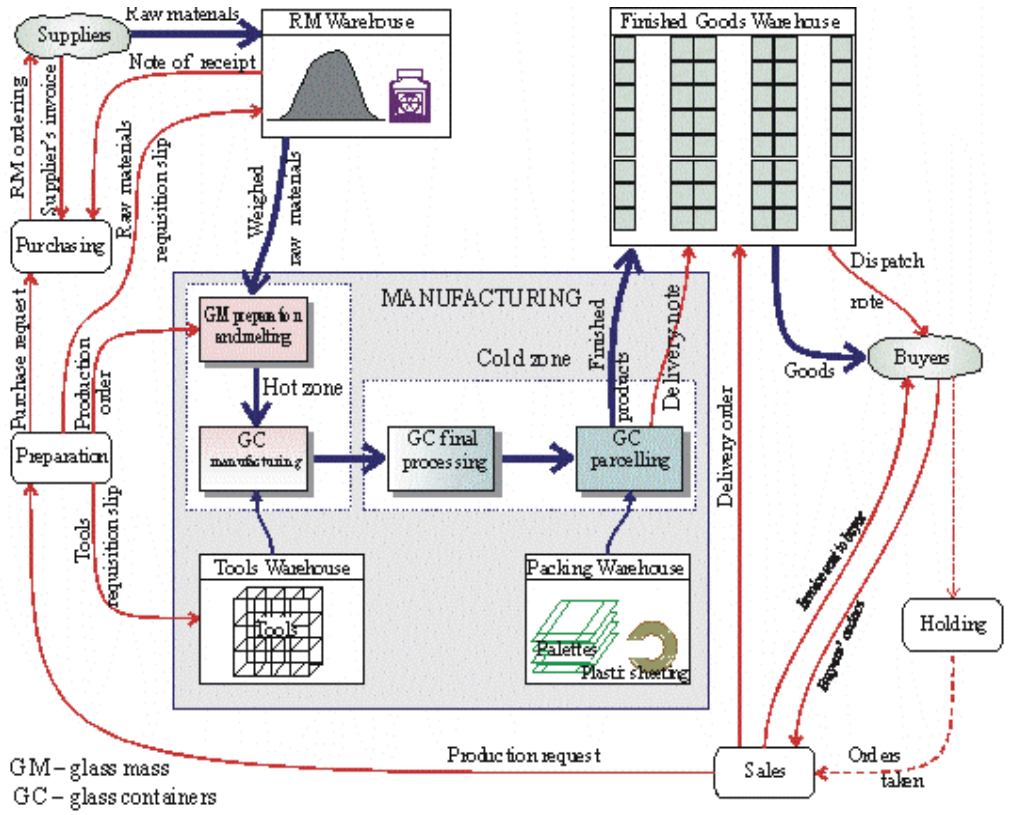


Fig. 2. Functioning of a glass container manuf. system's supply chain (Brumec et all, 1998)

In the raw materials warehouse materials and chemicals needed for uninhibited and continuous production of glass containers are stored. The renewal of supplies is conducted in accordance with the procurement plan and specific manufacturing requirements arising due to custom-made orders. A shortage of any raw material in the warehouse is not allowed as it may cause an interruption of production, which leads to closing down of the production line.

Tools and packaging warehouses are constituent parts of manufacturing since both tools and packaging are used in operations which are indispensable parts of the manufacturing process. The process of establishing the need for tools, as well as the ordering and receipt thereof is basically identical to that used in case of to raw materials. The need is defined by

manufacturing preparation after which purchasing occurs. The ordered tools and packaging are stored separately, with the manufacturing preparation responsible for supplies inventory.

After the final manufacturing stage finished products are delivered to the finished goods warehouse. In repetitive manufacturing guided by buyers' orders, a finished goods warehouse represents a buffer between the contradictory requirements for continuity of the manufacturing process on one hand and virtually unpredictable buyers' needs on the other. Production management is therefore equally based on buyers' individual orders and warehouse inventory. Although the finished goods warehouse is the sole responsibility of the sales function, it is closely connected with manufacturing preparation in matters concerning the planning of the quantity and type of products as well as production deadlines. Managing a finished goods warehouse is a complex task since such goods normally require large storage space, tend to be fragile and become obsolescent soon. As a result, taking finished goods inventory needs to include both quantitative stock-taking of individual items and inventory of batches (by production date, prescriptions and tools used) taking into consideration their physical location within the warehouse.

Regarding the complexity of managing a manufacturing business system with a combined continuous and discrete manufacturing process, it is necessary to point out the following features of such systems:

- Although the entire product range is manufactured from glass mass of varying features, it can be said that, according to their composition and colour structure, there is a relatively small number of glass mass types.
- Glass mass documentation (containing information about its composition, process parameters and procedures order in the hot zone) has been developed as a set of prescriptions, whereas tools required for manufacturing different glass container shapes have been described by means of construction design and a document specifying their components.
- Production documentation incorporates the prescriptions and the production tools to be used for manufacturing the product.
- Melted glass mass is a semi-product that several different types of products arise from in a continuous flow. For technological reasons, the melted glass mass cannot be stored for later finalization.
- In case of demand for a new type of glass (which seldom occurs) new prescriptions are first developed and tested, upon which the appropriate tools for forming glass containers are developed and tested; in case of demand for a new type of glass containers (which frequently occurs) new tools are designed, constructed and tested. Tool modifications are 'easier' than those in a glass mass type as they require less time and turn out to be more profitable.
- Production resources are synchronized with the technological operations sequence so as to continuously produce various types of glass containers from the same glass mix, using different tools as long as the planned production quantity does not require a modification of the glass mix.
- The principal goal of the logistic chain is to timely ensure the quantity of raw materials needed for a particular type of glass mass, reduce the frequency and duration of production delays caused by a change of tools or glass mass type, and guarantee a high reliability of the technical production system.

The major reference value in production management is the product quantity and the time by which the product ought to have been finalized and made available.

4. Simulation modelling of manufacturing business systems

Using simulations in a business environment is determined by the goals of a business system. If we assume that the general goal of each business system is its survival in the market and business growth, as well as the development of the system, then simulations will be directed at improving the execution of business processes, attempting to forecast future situations and identify critical success factors for achieving the desired business goals (Visawan&Tannock, 2004). In literature on simulation various authors propose different definitions of the concept of simulation, which depends on the area covered in a particular book or paper or a specific field of interest of the author. While some of them interpret simulation as a representation of systems dynamics, a mathematically grounded numerical technique, experimentation or computer software operations, others define it as a process or a process modelling technique. In general, it can be said that simulation modelling is mimicking of a real system by means of scientific methods (probability theory, statistics and operations research) and contemporary information technologies. It is in this way that the simulation process and its capabilities are most precisely defined.
Simulation modelling methods are a powerful means of achieving qualitative changes. Working with models makes it possible to view business systems dynamics, which means that individual action scenarios can be generated and the system fine-tuned in short time segments (Manzini et all, 2005). Each system state described by a conceptual model contains formally determined parameters and conditions which describe it. By changing the values of state indicators defined in such a way a range of possible future states of the system is obtained along with precisely defined border values, and conditions and ways of transition from one state to another(Ingemansson& Bolmsjö, 2004). This eventually implies that it is also possible to implement less radical yet continuous changes in the course of a business process, predict future events, and plan the procedures accordingly. Based on them, the system of management and monitoring of possible alternative states is developed that can serve as a procedural pattern for manufacturing business system management. By unification and formalization of manufacturing business processes, and their upgrade with a view to develop a meta-model, it is possible to create a general model of business systems states management(Lau &Mak, 2004). If input in the form of concrete values of individual system parameters is provided, such a model can be applied to a concrete business system within the class for which a given meta-model has been generated.

4.1 Conceptual models

Simulation modelling of a manufacturing business system starts with the development of the system's model. Models are based on the analysis of a business system's elements, structure, relations and functions. Well-developed models provide useful information on problem-solving possibilities as well as on alternative solutions, possible negative effects and states in which a real system can be found if certain system parameters are modified. Contemporary modelling presupposes the application of computers in defining the requirements and model construction. Nevertheless, the entire modelling process is still based on the knowledge, logic, ability of abstraction and experience of the individual that

develops the model so that the process cannot be fully automated. Using computers in the modelling process primarily refers to mathematical calculations of the value of particular parameters, since the complexity and the amount of such calculations grow exponentially depending on the increase in the number of elements that constitute a model. The first step in simulation modelling is the design of the system's conceptual model. Its purpose is to enable the structuring of the problem and its better comprehension. Conceptual models are important as they are meant to (Wenbinet all, 2006):

- isolate the essential characteristics of a system
- describe elements of the system and their interaction
- facilitate communication between the developers' team and model users
- assist in the computer model development.

Conceptual models contain a rough description of a system and its elaboration into separate modules. They represent a link between the problem-solving idea and the mental model of a real system on the one hand and strictly defined computer models that enable simulation of a system's behaviour on the other. There is a certain inaccuracy and inconsistency in representing a real system by means of conceputal modelling that originates from the inability to precisely determine the system's dynamic behaviour in time and possible occurrences of certain parallel activities in the system. Therefore a conceptual model is partly a general model, although one that can be viewed as a dynamic system in which objects of the model can also operate in parallel. Owing to these particular features, Petri nets are proposed as a suitable method for developing a conceptual business system model.

Petri nets, as a conceptual system modelling method, represent a means of modelling, researching and simulation of complex dynamic systems. They are used to predict the system behaviour and simulate future system states, besides determining the conditions of changes of system states. In that sense Petri nets are a special class of conceptual models which can be used for observing both current and future events, their sequence and conditions required for their occurrence and continuation (Wang et all, 2005).

In the observed glass container manufacturing system two separate segments can be recognized in the production function (Fig. 2) earlier referred to as the hot zone and the cold zone. In simulation modelling they can be represented as separate models that in a later stage of the simulation experiment will serve as the basis for determining the simulation type. Namely, as the production processes in the hot zone are performed continuously, the experiment will also be conducted by continuous simulation. On the other hand, the nature of the processes in the cold zone is discontinuous so discrete simulation proves a logical choice for simulating such a system. In modelling the hot zone the principal feature of that part of the production process needs to be considered, i.e. feedback system modelling. In doing so, methods and techniques of conceptual modelling in the domain of systems dynamics can be used that describe interactions between certain elements of the observed system. Figure 3 shows the model of hot zone functioning with feedbacks.

At the conceptual level it is possible to implement Petri nets for developing models of both zones provided that specificities of each of them are observed. In addition, Petri nets, as a fairly complex conceptual model, solve the inaccuracy inherent in causal loops. Petri nets show possible resources conflicts and the way in which they are resolved. Also, they highlight solutions for deadlocks, states conversion resulting from conflict resolution, and similar adverse system states. Petri nets (especially coloured Petri nets) fairly precisely describe system states conditions (continuous process with feedback), including a process network net, that is, the way of performing processes and states conversion.

The following entities are included in the hot zone: raw materials, mixer, mixture, furnace, monitoring and control system, and glass mass. Each of these entities is modelled by a Petri net, all of which are subsequently integrated into a model representing the entire system. Figure 4 shows the conceptual model of the control system which manages production processes in the hot zone in an automated manner. It ensures the continuous execution of the process by initiating certain actions based on the fulfillment of conditions.

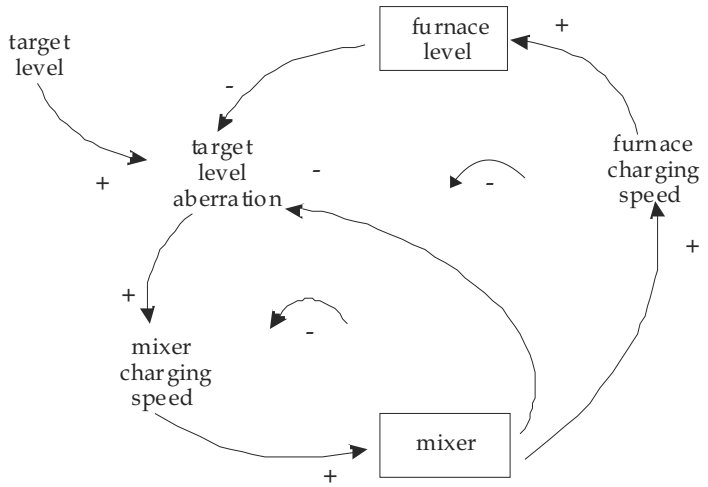


Fig. 3. Conceptual model of a causal loop

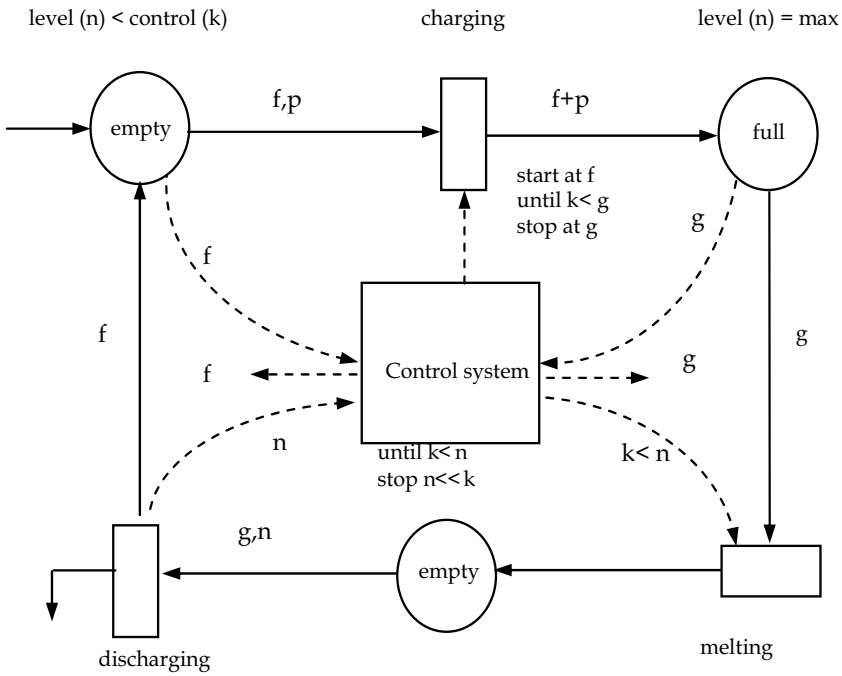


Fig. 4. Coloured Petri net of the hot zone control system

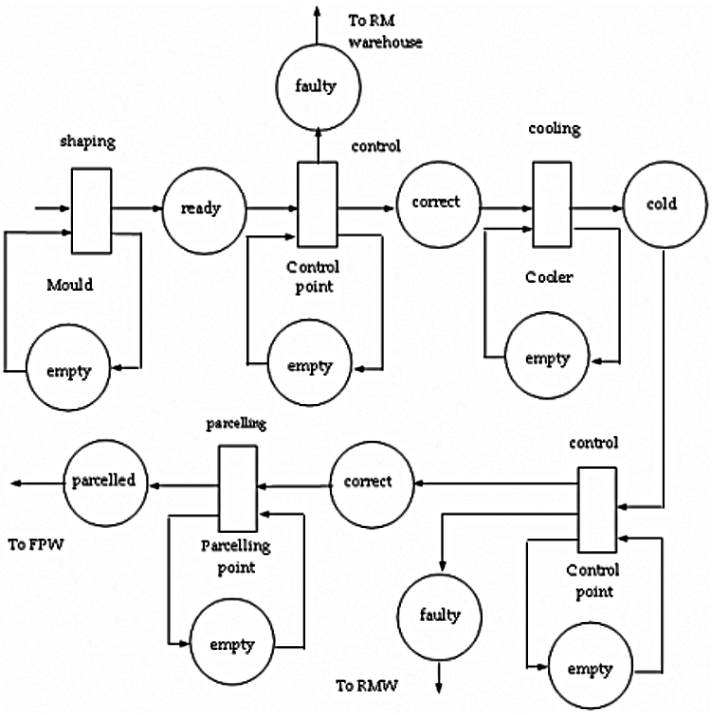


Fig. 5. Cold zone Petri net

The cold zone is the segment of the glass manufacturing process that comprises the final processing of glass containers until the moment when they are dispatched to the finished goods warehouse. In this segment of production glass containers undergo several subsequent processing stages before they are dispatched to the warehouse during which their characteristics remain unchanged.

This part of the manufacturing process is characterized by processes occurring on the assembly line. In other words, the assembly line conveys objects of processing from one processing site to another at which a specific action is performed. Apart from glass containers as the transitory entity the conceptual model of the hot zone also consists of entities that represent service points (production resources).

The process of conceptual model development ends with model validation that represents the verification of the correct logical functioning of conceptual models with regard to the real system. Validation is usually repeated during model adjustments until the desired acceptance level of the model is achieved. It is important to emphasize that by validation it cannot be determined whether a certain conceptual model fully corresponds to the real system. Instead, validation can establish that a certain level of equivalence with the real system is achieved.

4.2 Simulation models

Knowledge of a system's characteristics upon which a model is developed is not sufficient for successful simulation execution. Each real system, apart from its constituent elements and relations between them, has its function, or the way in which the elements and relations between them interact. Owing to this systems dynamics is achieved, which leads to creation

and exchange of certain system states. This means that in the process of simulation modelling a system of a model's behaviour is required along with the system's model itself, including the rules and algorithms of interactions among elements of the real system represented by the model(Kunnathuret all, 2004). Development of a computer simulation model of the hot zone is based on a set of business system features that are defined and described by the conceptual model. Conditions and assumptions from the conceptual model for the observed class of manufacturing business systems that execute part of their business processes in a continuous manner determine the implementation of systems dynamics as an appropriate method for the description of continuous business processes that include feedback. Feedback relationships among hot zone attributes are shown in the regulatory cycle in Figure 6. It shows systems dynamics attributes of entities in the hot zone as well as the direction in which they operate.

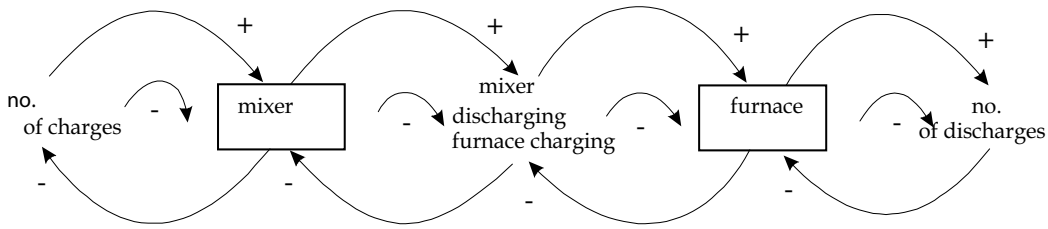


Fig. 6. Causal loop diagram of the process of glass mass preparation and melting

Systems dynamics that describes feedback systems is characterized by three key features:

- *level* – refers to the state of a resource or its accumulation;
- *rate* – refers to the rate of transforming resources from one state to another, that is, to level change rate;
- *delay* – refers to the time needed by the system to respond to the initiated action.

If the delay is shorter, the system can operate in a balanced way without drastic changes in behaviour and functioning. On the other hand, in case of a longer delay, oscillations arise in the functioning of the system (Doloi&Jaafari, 2002).

Features of hot zone systems dynamics and manufacturing processes therefore determine the manner in which a simulation model will be developed by a set of difference or differential equations that describe the states and conditions of changes of hot zone states. After the description of the system by means of a simulation model, programming is used to develop an application that will perform the simulation experiment on a computer.

By performing the simulation experiment data is obtained that demonstrate the way of a particular business system's functioning. The experiment is first conducted with data gathered in the real system to verify the computer model and the simulation program. Verification is the process of comparing the conceptual model with the computer model with a view to determine the equivalence in the functioning logic and certain parameters between the two models. The aim of this step is to establish whether all the important features of the system in the conceptual model are sufficiently well translated into the code of the computer simulation programme. Only after the validation and verification have been successfully performed is it possible to determine whether the results of simulation experiment will be sufficiently accurate and reliable to serve as the basis for making appropriate business decisions. However, apart from allowing for model verification, the

comparison of results and analysis of simulation results confirmed that the system which controls furnace recharging does not operate under optimal conditions.

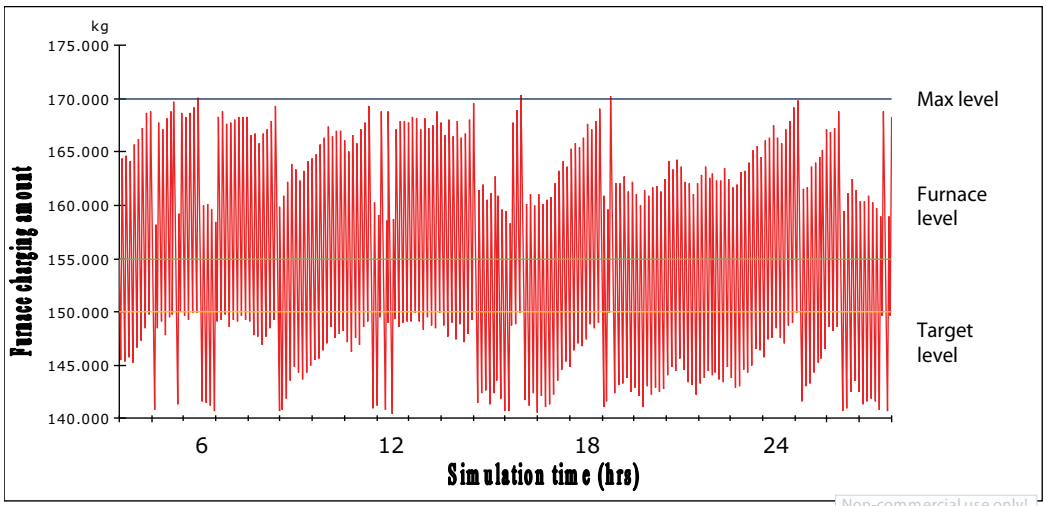


Fig. 7. Simulated furnace level state

Simulation experiments of the hot zone showed that the control system for the management of furnace operation cannot simply maintain the target level of the melted glass mass in the furnace at the currently defined parameters. In Figure 7 significant oscillations in the mean values of the level state and their aberration from the desired value are evident. In a real system this means that the manner of functioning of the control system is a stochastic function. In other words, the control system performs its activity, but the patterns of its functioning are not easily predictable.

Constant aberrations from the target level lead to the initiation of the furnace recharge cycle. Their total number is within the predefined values, but the duration of the cycle is long. As furnace discharging is executed continuously, and the glass mass level oscillates around the target level, a large number of level states assume values close to the minimum or even below it, which is potentially harmful as it can lead to system interruptions. Under normal circumstances they will not occur, but it still represents a permanent load that initiates regulation mechanisms. The control system for the management of furnace operation thus continuously generates control signals for process monitoring that should result in the desired state, but the system responds slowly and rarely achieves the desired state. Furthermore, the control system functions in a stochastic environment, without predictable functioning patterns, which increases the complexity of its functioning (i.e. the number of possible systems states and, consequently, patterns of functioning, is increased, which leads to a more complex decision-making process).

Such a scenario is certainly not desirable if a general states management model is to be defined for a particular type of business processes. The general meta-model of management should be simple, with a (relatively) small number of possible states, an easily predictable pattern of behaviour and clear conditions for state transitions. Therefore it is necessary to first implement certain changes in the model that will result in its better (that is, safer and

simpler) functioning. It is only then that the meta-model of states management should be defined.

The cause for such functioning of a management control system is the fact that the system's response to control signals is very slow. Research into technological processes has shown that the delay amounts to 15 minutes. The analysis of a conceptual model developed by a Petri net showed that this delay can be reduced. Figure 8 shows the proposed redesign aimed at shortening the delay.

In the real system model, after the signal indicating that the minimum furnace level has been reached is generated, the process of furnace recharging is initiated. The process starts at the raw materials warehouse. For all the actions in the technological procedure to be carried out, 15 minutes are needed. If the control signal of the level state is directed to the mixer by feedback at the moment when the mixer is already charged, the system's delay is reduced and amounts to only 7 minutes.

The duration of the delay cannot be further reduced as this time is used in the business system for technological actions between the mixer and the furnace and includes certain transportation activity as well as preparatory actions for furnace recharging (preparation and charging of furnace pre-chambers, temporary interruption in the melting process etc.). It should be taken into consideration that the entire manufacturing process is continuous, and the level to which the mixer is charged is the consequence of a previous state that can defined as the initial state at a given moment.

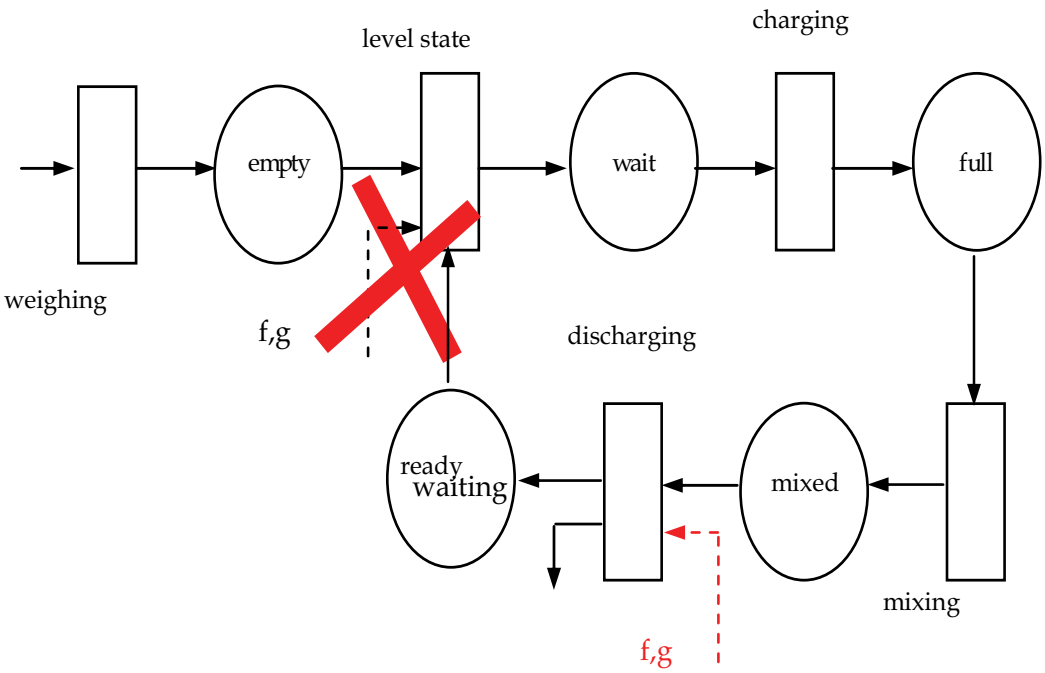


Fig. 8. Redesigned management

The currently observed state includes peak furnace load, with three production lines and different products, which requires a change in defining the state of the mixer level. Namely, the mixer level was previously defined as:

init 0 kg
Mixer level = +()dt – (Furnace charging rate)dt
d(Mixer level)/dt = Mixer charging rate (t) – Furnace charging rate (t)

The dynamics of mixer charging and discharging remains unchanged, but the definition of the initial state changes as well as the definition of the level itself. As a result, in the changed system the mixer level state is defined as:

init 500 kg
Mixer level = +(Mixer charging rate)dt – (Furnace charging rate)dt
d(Mixer level)/dt = Mixer charging rate (t) – Furnace charging rate (t)

In this way the condition which states that the mixer should always contain a sufficient amount of prepared mass regardless of the type of product manufactured on a particular line is fulfilled. In this model this value is constant for each instance of charging. As it is common that in the system the mixer is discharged entirely every time, this manner of changing the mixer level is maintained in the modified model as well.The proposed change does not require technological changes in the manufacturing process, and only refers to changes in the reverse cycle of information.However, the implemented changes entail major changes in the control function of the system for the management of furnace operation. Figure 9 shows simulation results for the redesigned model. It reveals that the changes introduced in the reverse cycle, which result in a reduction of delay, positively affect the functioning of the control system for the management of furnace operation. Balanced functioning of the control system is thus achieved as changes in furnace level states have a continuous flow in a particular direction.

Changes result in visibility of particular states of the observed system (i.e. processes related to discharging and recharging states can be monitored) and make it possible to more clearly define the conditions of transition of one state into another. This clear and continuous sequence of states is established as a default, which increases the predictability of the next state and simplifies states management. In other words, the control management system generates the identical type of control signals, in the identical sequence until control signals for the change of the sequence occur, which in turn results in the identical direction of the state sequence until the control state is reached. Such simple functioning can be generalized and described as a self-regulating management system for identical types of manufacturing processes. Moreover, it should not be disregarded that that such a manner of management can also be derived by means of mechanical dynamic automatic devices so that it is not necessary to use information systems and ICT solutions for system states management and control.

The simulation model of a production line comprises a set of manufacturing operations for the production of each particular glass container, from the very beginning of the line to product paletting as the terminal activity, and all the operations in between.

In the observed real business system this segment of production is performed continuously, with certain events occurring under the influence of the transactions that pass through them. Therefore it is possible to describe this part of the manufacturing system by a discontinuous simulation model. The production line of the observed business system is an assembly line that conveys glass containers from one processing location to another. At each point the processing capacity and time determine how long a certain glass container will be held at that location. The process is automated and monitoring and troubleshooting of possible delays on the assembly line is done by employees that supervise each particular processing location.

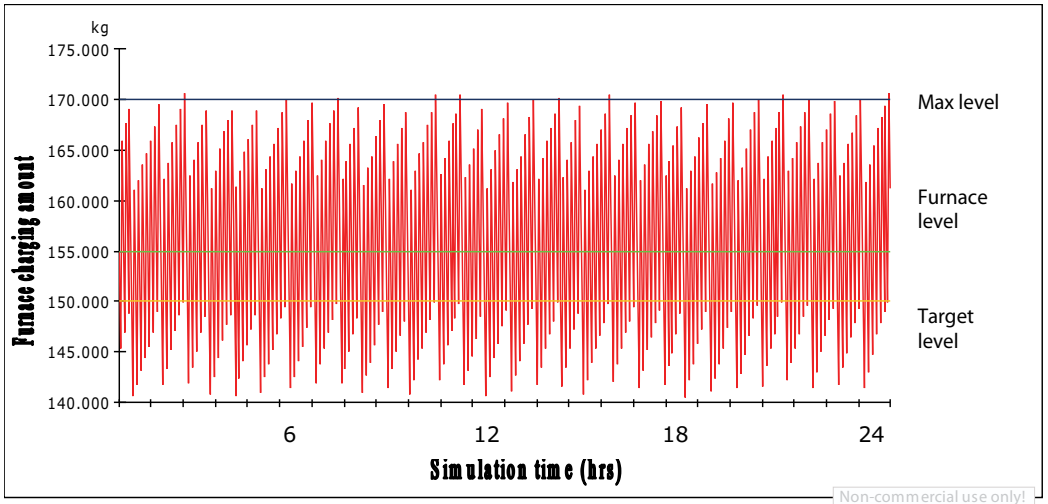


Fig. 9. Furnace level state after management redesign

The idea to be confirmed by means of simulation modelling is to investigate the possibility of manufacturing various glass container types by using a disposable number of production lines (that is, three production lines connected to the furnace). The production employing three production lines amounts to peak cold zone load. Possible combinations of simultaneous manufacturing of products in production lines whereby technological conditions of the manufacturing process are fulfilled need to be explored.

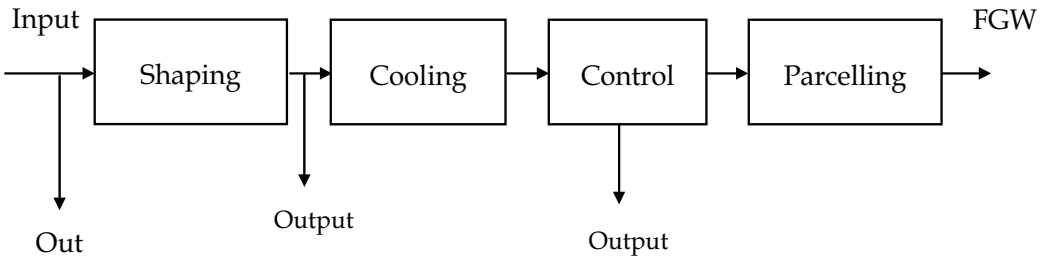


Fig. 10. Production line of glass container manufacturing

The principal condition to be observed in the model comprising three production lines is the maximum utilization of the melted glass mass, paralleled with the maximum utilization of production capacities. This means that production resources of production lines need to operate at the peak utilization degree, the prerequisite for which is a sufficient quantity of the melted glass mass.

Results of simulation experiments (Table 1) showed that simultaneous manufacturing of various types of products in disposable lines for container manufacturing is possible. In this way it was also confirmed that improving production planning, in other words, the dynamics of managing the production by means of orders, is possible. Using a particular combination of product manufacturing can be partial or total, depending on the production needs, delivery deadlines, finished products inventory, and other external factors (e.g. availability of glass mass within the same group to which the observed business system belongs).

Model	No. of inputs by mould capacity	Amount by no. of inputs (kg)	No. of inputs used in the model	Amount by no. of inputs used (kg)	Amount disposable by the beta distribution (kg)	Difference real/sim (kg)
"AAB"	92160 + 40320	37232,64 + 19353,6	91979 + 40046	37159,51 + 19222,08	58.222,44	+1636,2 +1840,85
"AAC"	92160 + 43920	37323,46 + 19324,8	91979 + 44133	37159,51 + 19418,52		+1574,18 +1644,41
"ABB"	46080 + 80640	18616,32 + 38707,2	45990 + 80093	18579,96 + 38444,64		+898,92 +1197,84
"ACC"	46080 + 87840	18616,32 + 38649,6	45990 + 88265	18579,96 + 38836,6		+956,52 +805,88
"ABC"	46080 + 40320 + 43920	18616,32 + 19353,6 + 19324,8	45990 + 40046 + 44132	18579,96 + 19222,08 + 19418,08		+927,72 +1002,32
"BBB"	120960	58060.8	120139	57666.72		+161,64 +555,72
"BBC"	80640 + 43920	38707,2 + 19324,8	80093 + 44133	38444,64 + 19418,52		+190,44 +359,28
"BCC"	40320 + 87840	19353,6 + 38649,6	40046 + 88265	19222,08 + 38836,6		+219,24 +163,76
"CCC"	131760	57974.4	132398	58255.12		+248,04 - 32,68

Table 1. Overview of planned, disposable and simulated glass mass amount and product units

5. Conclusion

Researching manufacturing business systems does not only enable to collect measurable data on possible system states and parameters applicable inside and outside a business environment, but also to determine conditions and causes that lead to changes of business system states. Defining initial models by using Petri nets enhances the simulation modelling process since the defined models that are formally described at the conceptual level make the process of conducting experiments on simulation models more efficient. The validation of the defined models improves their trustworthiness, which in turn contributes to the certainty of the final results obtained by the simulation experiment. Verifying possible alternatives through simulation experiments and validating the results of those experiments improve the reliability and quality of future business system design. In this way the quality and reliability of decision making in the process of system design is also enhanced, as it is based on verified alternative models. An optimal business systems design mode is thus obtained, which eventually results in rational utilization of business resources.

6. References

Bider I. (2005). Choosing Approach to Business Process Modeling – Practical Perspective, Journal of Conceptual Modeling, Issue 34

Brumec J.; Dušak V. &Vrček N.(1998). Informacijski sustavi poduzeća VETROPACK STRAŽA d.d. – strateški plan razvoja, osnovni projekt i izvedbeni zahtjevi, Varaždin-Hum na Sutli

Dewhurst, F.; Barber. K. & Rogers, J.J.B. (2001). Towards integrated manufacturing planning with common tool and information sets, International Journal of Operations & Production Management, Vol 21 No 11, pp. 1460-1482.

Dilworth B. J.(1999). Operations Management – Providing value in Goods and Services, The Dryden Press Orlando

Doloi H. &Jaafari A. (2002).Conceptual simulation model for strategic decision evaluation in project management, Logistics Information ManagementVolume 15 Number 2, pp. 88-104

Ingemansson, A. & Bolmsjö, G.S.(2004). Improved efficiency with production disturbancereduction in manufacturing systems based on discrete-event simulation, Journal of Manufacturing Technology Management, Vol 15 No 3, pp. 267-279.

Jiming L.; XiaoLong J. & Kwok C.T. (2004). Autonomy Oriented Computing: From ProblemSolving to Complex Systems Modeling, Springer

Kunnathur A. S.; Sundararaghavan P. S. &Sampath S. (2004). Dynamic rescheduling using asimulation-based expert system, Journal of Manufacturing Technology Management; Vol 15 No 2, pp. 199-212

Lau, H.Y.K. &Mak, K.L. (2004). The design of flexible manufacturing systems using an extended unified framework, Journal of Manufacturing Technology Management, Vol 15, No 3, pp. 222-238

Manzini, R.; Ferari, E.; Gamberi, M.; Persona, A. & Rigattieri, A. (2005). Simulation performance in the optimisation of the supply chain, Journal of Manufacturing Technology Management, Vol 16 No 2, pp. 127-144.

Visawan D. &Tannock J. (2004). Simulation of the economics of quality improvement in manufacturing. International Journal of Quality & Reliability Management; Vol 21 No 6, pp. 638-654

Wang, Z.; Zhang, J. & Chan, F.T.S. (2005). A hybrid Petri nets model of networked manufacturing systems and its control system architecture, Journal of Manufacturing Technology Management, Vol 16 No 1, pp. 36-52.

Wenbin Z.; Juanqi Y.; Dengze M. & Ye J., Xiumin F. (2006). Production engineering-oriented virtual factory: a planning cell-based approach to manufacturing system design. International Journal of Advantage Manufacturing Technology; Vol 28, 957-965

Zhao F.; Hong Y.; Yu D.; Zhang Q. & Yi H.(2007). A hybrid algorithm based on particle swarm optimization and simulated annealing to holon task allocation for holonic manufacturing system, International Journal of Advantage Manufacturing Technology; Vol 32, 1021-1032

6

Project-Driven Concurrent Product and Processes Development

Janez Kušar, Lidija Rihar, Tomaž Berlec and Marko Starbek
University of Ljubljana,
Faculty of Mechanical Engineering
Slovenia

1. Introduction

When entering the global market, companies encounter several difficulties, the most severe being long product development times and too high costs of sequential product and process development. In order to overcome this problem, the companies will have to make a shift from sequential product and processes development (which is wasteful regarding time and costs) to a project-driven concurrent product and processes development as soon as possible.

"Customer is the king!" is becoming the motto of the global market. In the competition between suppliers of products only those companies will survive, which can offer innovative and individual products of good quality, produced in shortest possible time and at the lowest price (Eversheim et al., 1995).

Strong competition, existence of the market of customers and increased complexity of products and processes are the characteristics of today's competition.

Fast product and process development, combined with timely participation of customers and suppliers, together with entering the market at the right time, seem to be the decisive criteria for the market success of a product. The first supplier of a new product on the market has an advantage over the competition and thus he has the possibility of a faster return of product development investments (Duhovnik et al., 2001).

The company has to switch from sequential to concurrent product and process development (i.e. from sequential to concurrent engineering) in order to reduce product and process development time, reduce development costs and ensure quality of the product according to the customer's wishes (Prasad, 1996).

The paper presents a procedure for project-driven concurrent product and processes development by taking into account three strategic management methods: parallelness, standardisation and integration of product development processes. Also presented are the changes in organisational concept of the company, organisation of processes, organisation of work and organisation of IT, which are required for a transition from sequential to concurrent product and processes development.

Finally, an analysis is presented on concurrent product and processes development teams in a company; this analysis is a prerequisite for a transition to a new method of product and processes development.

2. Integration of project management and concurrent product development

The company that chooses project management of concurrent product and processes development first has to make project management system guidelines: rules of procedure, project management manual and operative instructions for project management, which precisely describe the implementation procedure of project phases from the bid to the end, as well as evaluation of the project.

For each concurrent product and processes development project it is necessary to set up a project dossier – a data warehouse of all data produced in the project lifetime. The project dossier has to be accessible to all project participants via the Internet.

According to the research (Kušar et al., 2008), the process model of project management of concurrent product and processes development has to contain logical sequence of product and processes development project activities and documents that arise from execution of activities (Figure 1).

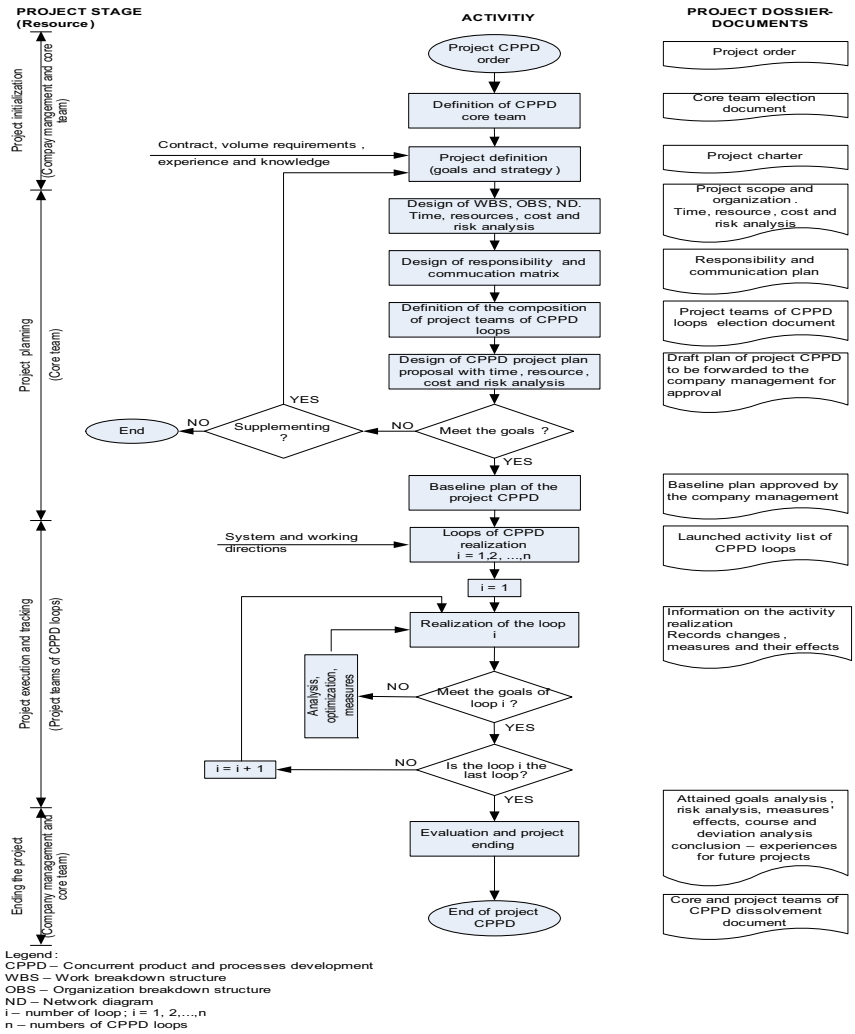


Fig. 1. Process model of project-driven concurrent product and processes development

It can be seen from Figure 1 that the process model of project management of concurrent product and processes development consists of four steps:

Step 1. Definition of objective of project management of concurrent product and processes development – order and definition of the project.

Step 2. Planning the concurrent product and processes development project: planning the WBS / project structure, organisation of project implementation OBS, responsibility matrix, network diagram and basic project activity plan.

Step 3. Execution and monitoring of the concurrent product and processes development project - project manager (via the project management office) takes care of project activity implementation.

Step 4. Completion of the project - when the project has been completed, evaluation is made, including analysis of the results achieved.

2.1 Strategic management during project-driven concurrent product and process development

Prerequisite for a successful project management of concurrent product development requires three levels of strategic management (Bullinger & Warnecke, 1996), i.e. parallelness, standardisation and integration of product development processes, as shown in Figure 2.

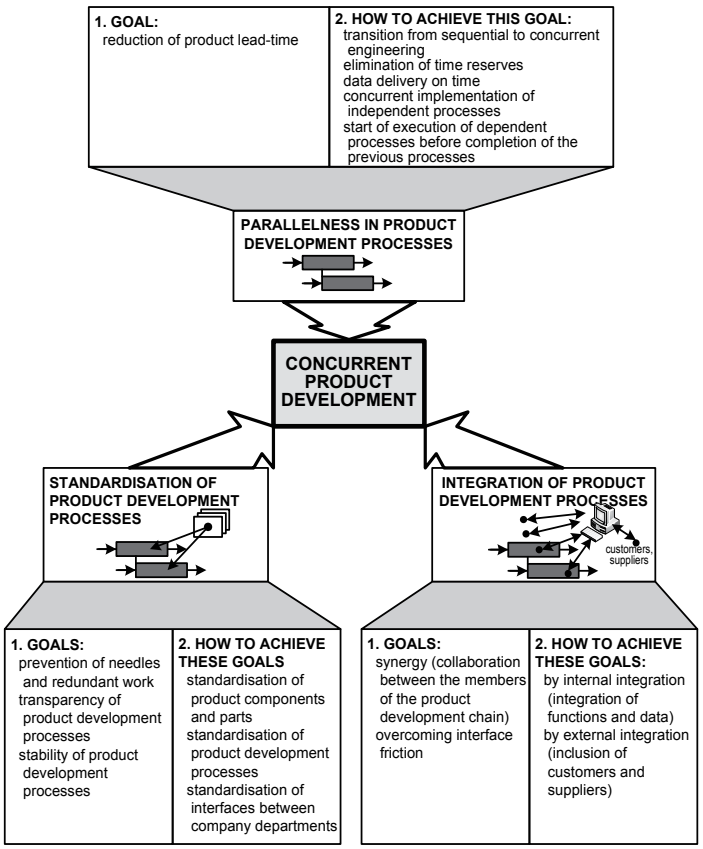


Fig. 2. Strategic management during project-driven concurrent product and processes development

The product and processes development time is reduced considerably by parallelising new product development processes. Independent processes, which are executed one after another in sequential development, are executed parallelly during concurrent development. In concurrent product and processes development the execution of interdependent product development processes starts before the previous processes have been completed, and thus the portion of uncertain and uncompleted data increases.
An advantage of parallel product development processes is fast execution of networked processes, while a disadvantage is increased transfer of data between product-development teams.
Standardisation of product development processes means the description and management of various views on product development processes, which is continuous and independent of individuals and events.
Standardisation applies to: product components (modules, components, parts), processes for manufacturing product components, and organisational plan for the implementation of product components (interfaces between departments, project approach).
By standardising product development processes, redundant and unnecessary work is avoided, higher transparency and stability of processes is achieved and thus more time for execution of innovative and creative tasks is ensured.
All company departments, as well as customers and suppliers should be part of the chain creating the features of the product under development. However, this leads to high interface losses because of uncoordinated scheduling, various interpretations of the roles of tasks and unknown requirements of internal customers.
Integration with direct inclusion of all company departments, customers and suppliers into the product development processes allows for a possibility of overcoming collisions at interfaces. Interdisciplinary work, process-oriented thinking and functioning, as well as creativity and conscious decision-making require integrated product development processes.
The goal of product development process integration is therefore a transformation of separate interfaces into a coherent whole.

2.2 Changes in the company before transition from sequential to concurrent engineering

According to our experience in the field of project management and concurrent product and processes development, a company that wants to make a transition from sequential to project-driven concurrent product and processes development should change the organisational concept of the company, organisation of processes, organisation of work and organisation of IT.
Organisational concept of the company defines the structure and competences of employees, who will be engaged in the concurrent product development – it therefore defines the mode of organisational unit formation and coordination between them.
In sequential product development there is a precisely defined hierarchy of reporting, as well as the procedure and competences of decision-making. Concurrent product and processes development requires a project management of product development process and a transition from individual- to team-work.
The basis of team-work is cooperation between team members and their interdependence (successful communication between team members ensures the team success). Team-work is performed by team members. Their main tool is communication and none of the team members may leave the team until the work has been completed. Team-work is a form of

collaboration between team members who are responsible for the distribution and implementation of tasks, for the solution of problems, as well as for communication within the team and between teams.
Good organisation and implementation of team-work is essential for a successful transition from sequential to project-driven concurrent product and processes development.
Success of concurrent product and processes development depends largely on the planning and management of product development project. During planning and management, a special attention has to be paid to the standardisation of product development processes, information transfer methods and the fastest possible integration of all members of the product development team, including customers and suppliers.
Figure 3 presents changes in the organisational concept of the company that decided to make a transition from sequential to concurrent product development.

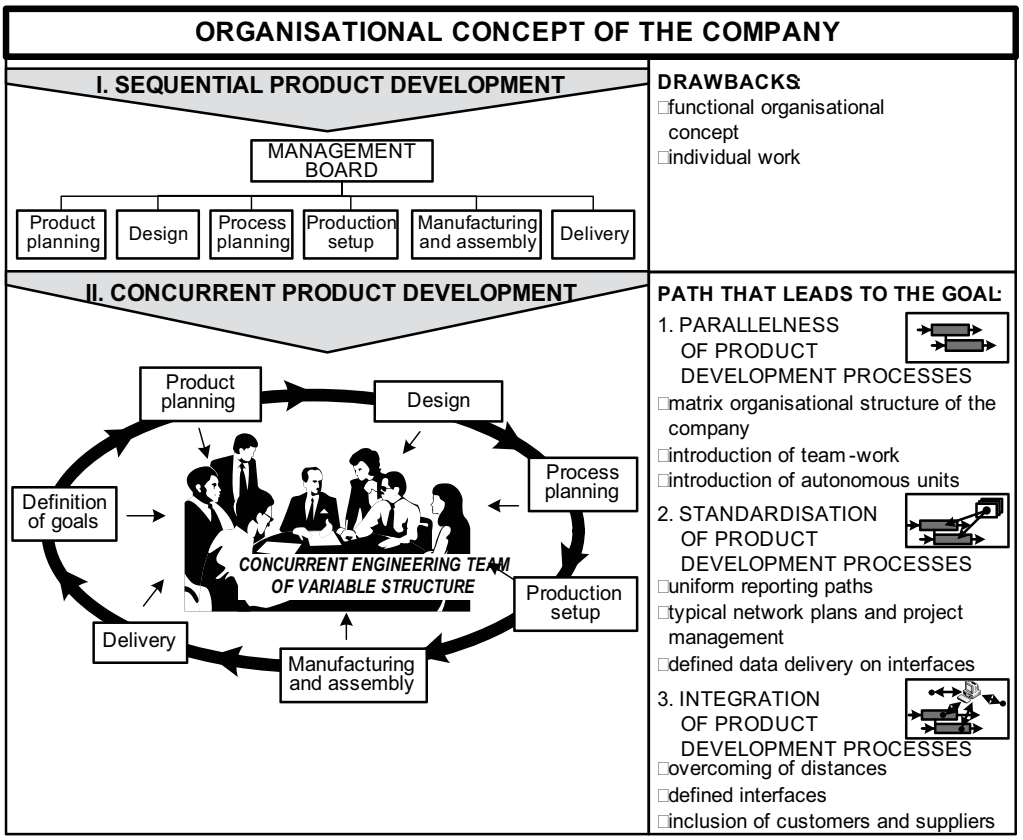


Fig. 3. Changes of the company organisational concept

A specific feature of product development processes is that they are goal- and result-oriented (Eversheim et al., 1995)
Analysis of the current sequential product development processes is a basis for planning concurrent product development processes. The barriers between company departments and between the company and its customers and suppliers can be eliminated if the company mind structures are changed. Standardised process descriptions are essential for parallel execution of product-development-process activities.

All product-development-process data should be prepared in the same way, so that during development of the planned product the results of the previously developed products can be used.

Figure 4 presents changes in organisation of processes in the transition from sequential to concurrent product and processes development.

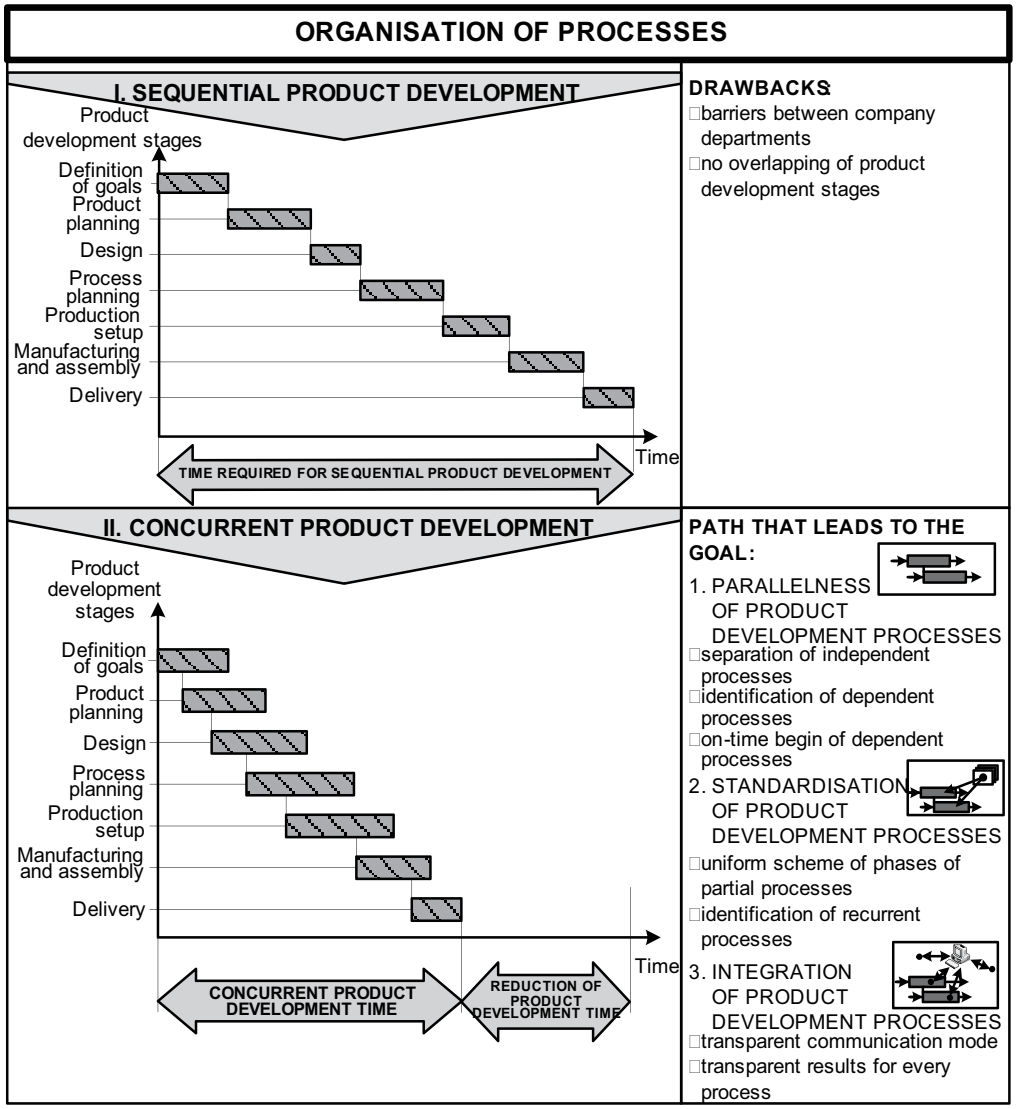


Fig. 4. Changes in the organisation of processes

A company planning to make a transition from sequential to project-driven concurrent product and processes development needs employees capable of team-work and rotation of work.

Team-work is successful when the team output exceeds the sum of outputs of team members, working individually.

According to the recommendations (Lencioni, 2002):

- small team consists of 2 to 25 members;
- large team consists of more than 25 members.

However based on experience of LAPS the upper limit for one team is 15 members.
Figure 5 shows team management methods with respect to suggested team sizes.

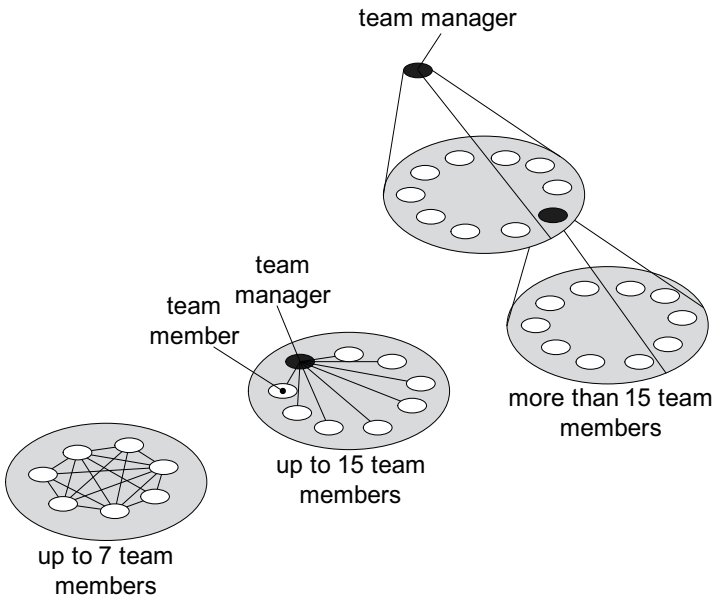


Fig. 5. Team management methods

Team of up to 7 members operates without an appointed team manager – team members manage themselves.
In a team of up to 15 members, a team manager is appointed; (s)he should be a team member.
If a team consists of more than 15 members, it is divided into several sub-teams in order to ensure successful team management. A team manager is not supposed to be a team member – (s)he should work only as a team manager.
If there are up to 7 members in a concurrent product and processes development team, each team member is interconnected with each other. Any team member can start a communication and all team members have the same decision rights. There is close collaboration between team members and they are satisfied with their work.
Capability of team-work means openness in sharing of information and admitting errors, as well as responsibility for decisions, so that individual tasks will be performed at the right time and thus the highest possible parallelness will be achieved.
Team members should have rotational-work capability, which is essential for understanding views of others on the problems encountered (Eversheim et al., 1995).
Figure 6 presents changes in the organisation of work in a transition from sequential to the project-driven concurrent product and processes development.
Transition from sequential to project-driven concurrent product and processes development requires large IT investments.
The data must be accessible on all of the product-development-process locations. This approach allows for faster transformation processes of the most important forwarded data (the push principle) or the data fetched from the databases (the pull principle).

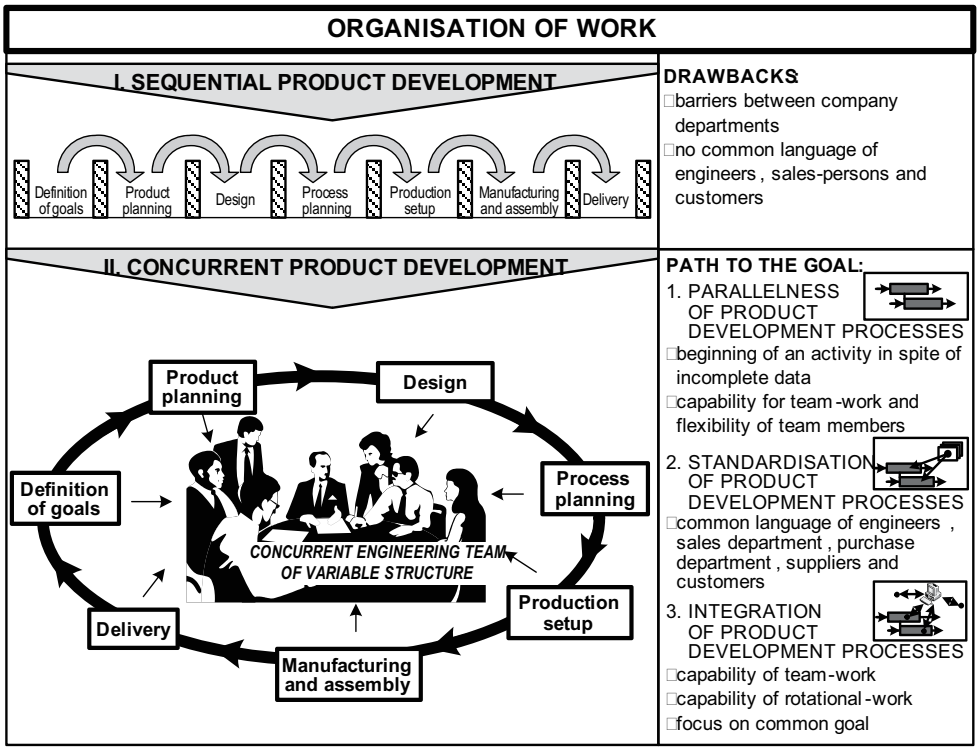


Fig. 6. Changes in the organisation of work

Changes of IT, required by concurrent product development, support parallelness, standardisation and integration of product development processes.
Figure 7 presents changes of IT in the transition from sequential to project-driven concurrent product development.

3. Case study: how is a company prepared for concurrent product and processes development

A company – a component developer and supplier for the automotive industry – decided to perform an analysis regarding fulfilment of the basic conditions for a transition from sequential to concurrent product and processes development.
For this purpose the company hired an external counsellor – the Laboratory for Manufacturing Systems (LAPS) of the Faculty of Mechanical Engineering, Ljubljana, Slovenia – to find out (together with the company management):

- whether the people who will be members of the project-driven concurrent product and processes development team are capable of and motivated for team-work, and whether any of the five team-dysfunctions exist in the team (Lancini, 2002),
- do team members have proper personal value systems for team-work (Ellis et al., 2006),
- are all nine team-roles represented in a team (Belbin, 2003)?

The company management decided that the analysis of the team-work efficiency would be tested in a team for development of a car pedal component (Figure 8). The team consists of eight members, as shown in Figure 9.

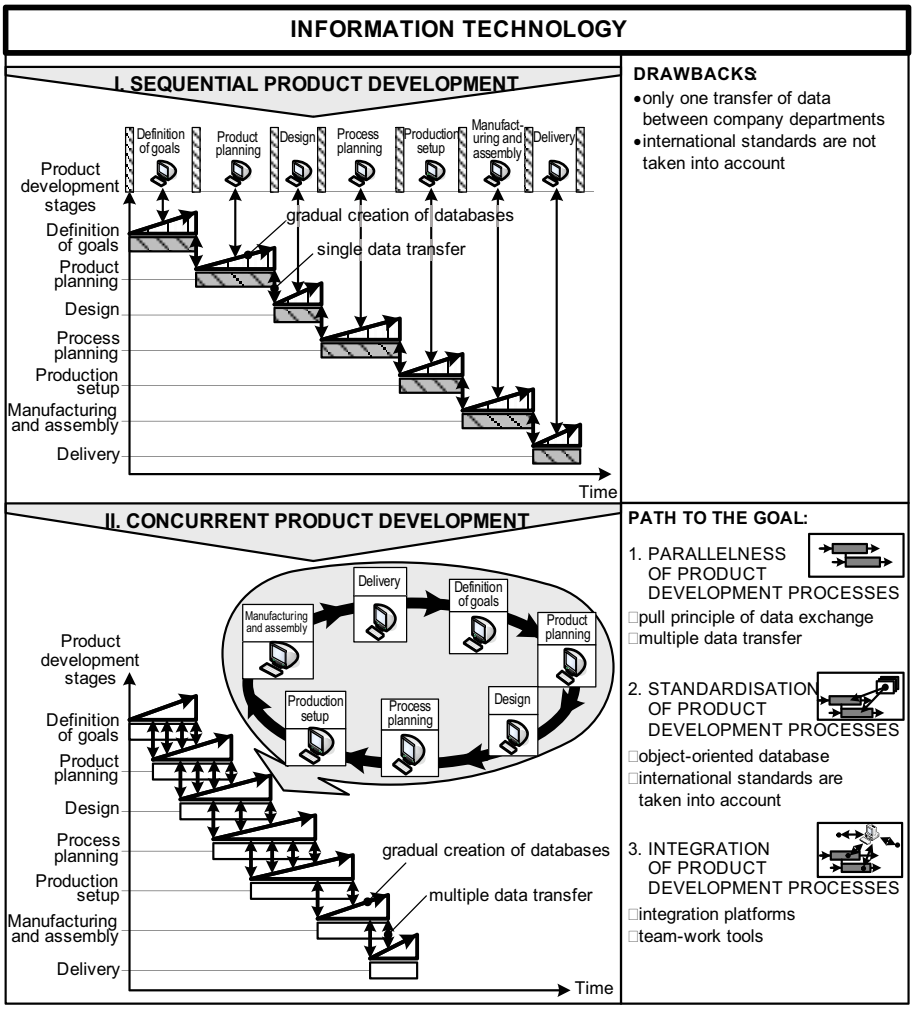


Fig. 7. Changes of IT

Fig. 8. Car pedal component

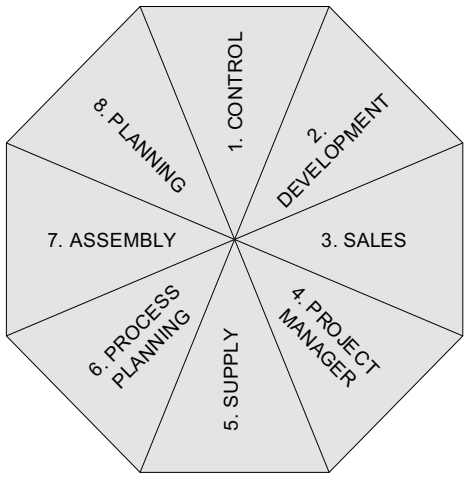


Fig. 9. Structure of the analysed team

3.1 Analysis of team-work capability, motivation and team dysfunctions

In order to find out whether team members who will in future participate in project-driven concurrent development of car pedal component are capable of team-work, the LAPS employees modified the questionnaire created by Lancini (Lancini, 2002) on team-work capabilities (shown in Table 1).

QUESTIONNAIRE 1: TEAM-WORK CAPABILITIES	
1. In a team – do you tend to appropriate to yourself the results of the work performed? …	_____
2. Do you complement other team members regarding knowledge and capabilities? ……	_____
3. Does the team-work facilitate development of communication skills? ……..	_____
4. Do team members help each other? ……..	_____
5. Is there a constructive criticism in the team? ……..……..	_____
6. Do all team members participate in decision-making? ……..	_____
7. Do team members support adopted decisions? ……..	_____
8. Does any individual team member get more support? ……	_____
9. Are the solutions of problems that arise in team-work assessed together and analysed critically? ……	_____
10. Are there many ideas collected regarding a problem solution? ……..	_____
11. Are the team members capable of adopting compromise solutions of the problem? ….	_____
12. Do conflicts arise between team members? ……	_____
13. Are team members who agree with proposals of the stronger members exposed to any pressure?	_____
14. Is the team under control of an individual? ……	_____
Allocate 1 to 3 points to each answer: 1 point – rarely true 2 points – sometimes true 3 points – usually true	
FINDINGS: • If the total number of points is between 33 and 42 → team member IS CAPABLE of team-work. • If the total number of points is between 24 and 32 → team member IS PARTIALLY CAPABLE of team-work. • If the total number of points is between 14 and 23 → team member IS NOT CAPABLE of team-work.	

Table 1. Team-work capability questionnaire

All eight team members answered to 14 questions regarding team-work capabilities. Results of the analysis are shown in Table 2.
Results in Table 2 indicate that four team members are capable of team-work, the other four ones are only partially capable, and there is no team member who would not be at least partially capable of team work.
Final conclusion: the team is capable of team-work.

No.	TEAM MEMBER	TOTAL NUMBER OF POINTS REGARDING TEAM-WORK CAPABILITY	TEAM WORK CAPABILITY
1.	Control	30	Partially capable
2.	Development	40	Capable
3.	Sales	41	Capable
4.	Project manager	41	Capable
5.	Supply	39	Capable
6.	Process planning	29	Partially capable
7.	Assembly	26	Partially capable
8.	Planning	31	Partially capable
	AVERAGE CAPABILITY	**34.6**	**CAPABLE**

Table 2. Team-work capability

In order to find out whether team members are motivated for team-work the questionnaire in Table 3 was compiled.

QUESTIONNAIRE 2: TEAM-WORK MOTIVATION	
1. Is your work useful?	_____
2. Do you know the purpose of your work?	_____
3. Do you know the results of your work?	_____
4. Do you have good working conditions?	_____
5. Does team manager praise you?	_____
6. Does team manager criticise you?	_____
7. Does team manager give you instructions for work?	_____
8. Do you compete with other team members?	_____
9. Do you actively participate in the team?	_____
10. Does team manager create problems on purpose?	_____
11. Does team manager set work objectives properly?	_____
12. Does team manager take care for solution of conflicts?	_____
13. Does team manager take care for a good team atmosphere?	_____
14. Are you motivated by your salary?	_____
Allocate 1 to 3 points to each answer: 1 point – rarely true 2 points – sometimes true 3 points – usually true	
FINDINGS: • If total number of points is between 33 and 42 → team member IS MOTIVATED for team-work. • If total number of points is between 24 and 32 → team member IS PARTIALLY MOTIVATED for team-work. • If total number of points is between 14 and 23 → team member IS NOT MOTIVATED for team-work.	

Table 3. Questionnaire on team-work motivation

All eight team members answered to 14 questions regarding team-work motivation. Results of the analysis of their answers are shown in Table 4.

No.	TEAM MEMBER	TOTAL NUMBER OF POINTS REGARDING TEAM-WORK MOTIVATION	TEAM WORK MOTIVATION
1.	Control	39	Motivated
2.	Development	40	Motivated
3.	Sales	30	Partially motivated
4.	Project manager	39	Motivated
5.	Supply	30	Partially motivated
6.	Proces planning	39	Motivated
7.	Assembly	27	Partially motivated
8.	Planning	36	Motivated
	AVERAGE MOTIVATION	**35**	**THE TEAM IS MOTIVATED**

Table 4. Team-work motivation

Results in Table 4 indicate that five team members are motivated for team-work while the other three members are only partially motivated.
Final conclusion: The team is motivated for team-work.
The following five dysfunctions can arise in teams (Lencioni, 2002):

- Dysfunction 1: Lack of confidence
- Dysfunction 2: Fear of conflicts
- Dysfunction 3: Non-interoperability
- Dysfunction 4: Rejection of responsibility
- Dysfunction 5: No interest for results.

In order to find out whether any dysfunction (or which dysfunction) exists in our team, the employees of the LAPS modified the questionnaire created by Lancini (Lancini, 2002) shown in Table 5.
All eight team members answered to 15 questions regarding susceptibility to team-dysfunctions. Answers of one of the team members are shown in Table 6.
Analysis of results of all eight team members indicated that the team's dysfunction is "no interest for results". In order to eliminate the sources for this dysfunction in future, it will be necessary to:

- precisely define objectives of team work,
- support behaviour and mode of operation that lead to the defined objectives.

3.2 Suitability of personal value systems for team-work

We used the SDI (Strength Deployment Inventory) method (Ellis et al., 2006) to analyse the personal value systems of eight team members.
This method can be used to find out:

- how an individual team member understands himself and his value system in relation with other team members and other project participants,
- how a team member understands other team members and how he takes this fact into consideration in mutual relations with them,
- how a team member reacts in a conflict situation and how he understands reaction of other team members in a conflict.

QUESTIONNAIRE 3: SUSCEPTIBILITY TO TEAM-DYSFUNCTIONS		
Dysfunction 1: LACK OF CONFIDENCE	1. Team members sincerely apologise for improper words or deeds which may harm the team operation.....	_____
	2. Team members admit their faults and weak points.....	_____
	3. Team members know each other personally and they often talk about their private lives.....	_____
Dysfunction 2: FEAR OF CONFLICTS	4. During discussions each team member openly and honesty says what he thinks.....	_____
	5. Team meetings are exciting and never boring.....	_____
	6. During meetings, not only important themes are being discussed, but also the most difficult ones that are being dealt with openly and common solutions are found.....	_____
Dysfunction 3: NON-INTEROPERABILITY	7. Team members know projects of their colleagues and know the contribution of each individual to the common goal of the team.....	_____
	8. After the meeting, the team members are sure that everyone completely agrees with adopted decisions, although they had different opinions at the beginning.....	_____
	9. Team members finish their discussions with clear and unambiguous results and plans.....	_____
Dysfunction 4: REJECTION OF RESPONSIBILITY	10. Team members remind each other on improper or unproductive behaviour.....	_____
	11. Team members are ready to support their colleagues.....	_____
	12. Team members ask each other on progress of work regarding the plan and their acts.....	_____
Dysfunction 5: NO INTEREST FOR RESULTS	13. Team members are ready to accept limitations in their departments or their fields of work (e.g. budget reduction, responsibilities, number of employees, etc.) if this is in favour of the team.....	_____
	14. Working enthusiasm suffers if common goals have not been met.....	_____
	15. Team members do not expect continuous praises for their achievements; instead they spontaneously praise good results of others.....	_____
Allocate 1 to 3 points to each answer: 1 point – rarely true 2 points – sometimes true 3 points – usually true		
FINDINGS: • If the total number of points for a particular dysfunction is between 8 and 9 → there is a high probability that this particular dysfunction does not cause a problem in the team. • If the total number of points for a particular dysfunction is between 6 and 7 → it is possible that this particular dysfunction may start causing problems in the team. • If the total number of points for a particular dysfunction is between 3 and 5 → the team has to fight against this particular dysfunction because it exists in the team.		

Table 5. Questionnaire on susceptibility to team-dysfunctions

DYSFUNCTION	Answer No.	Answer No.	Answer No.	TOTAL POINTS
Lack of confidence	1. *3*	2. *2*	3. *3*	8 - dysfunction exists
Fear of conflicts	4. *2*	5. *3*	6. *2*	7 - dysfunction may exist
No connections	7. *2*	8. *2*	9. *2*	6 - dysfunction may exist
Rejection of responsibility	10. *2*	11. *3*	12. *3*	8 - dysfunction does not exist
No interest for results	13. *2*	14. *1*	15. *1*	4 - dysfunction exists

Table 6. Susceptibility to team-dysfunctions – member from the DEVELOPMENT team

For implementation of the SDI method a workshop was organised which showed positions of team members in the SDI triangle according to their value systems. All eight team members answered to 20 questions regarding personal value systems in a team-work.
Their answers and their positions in the SDI triangle are shown in Figure 10.

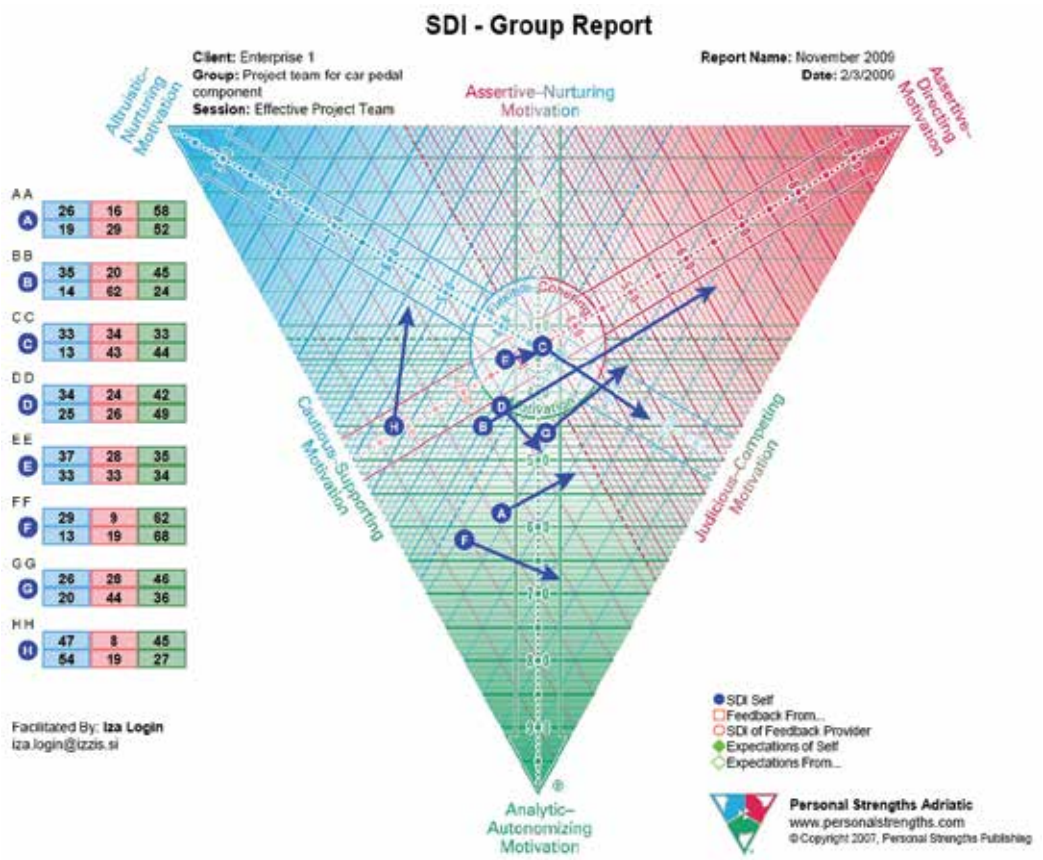


Fig. 10. Positions of team members in the SDI triangle

An arrow is assigned to each team member. The starting point of the arrow defines personal value systems of the team member in a non-conflict environment, while the arrow represents the mode of his/her reaction in various conflict stages.
Table 7 shows personal value systems of team members, their understanding of conflict situations and their suitability for team- and individual work.

Team member	Personal value system	Understanding of conflict situations	Suitability for team-work
1 Control	Analytic-autonomizing type	When encountering difficulties he first responds with analytical, logical and restraint approach, followed by an unyielding, strong attack, based on logic and planning. If these approaches do not help, he gives up for the sake of harmony, but he chooses this as the last resort.	He functions in the team, although individual work suits him more.
2 Development	Cautious-supporting type	He wants to achieve his domination by competition. If competition does not help, he uses analysis, logic, sense and rules. He selects withdrawal as the last resort.	Good feeling for team-work.
3 Sales	Flexible-cohering type	He does not respond directly to difficulties; instead he uses several strategies for solution. He intelligently works his way to reach the objective, but he gives up in a situation without prospects.	Strong feeling for team-work.
4 Project manager	On the border between flexible-cohering and cautious-supporting type	In difficulties he first takes the position based on logic, order, rules and principles. In next stage of conflict he chooses one of the following possibilities: if the case is important to him, he fights; otherwise he gives up.	Strong feeling for team-work.
5 Supply	Flexible-cohering type	His responses to difficulties and opposition vary considerably. His response depends on situation and circumstances and has no firm sequence.	Strong feeling for team-work.
6 Process planning	Analytic-autonomizing type	He wants to achieve his domination with competition. If competition does not help, he uses analysis, logic, sense and rules. He selects withdrawal as the last resort.	Good feeling for team-work.
7 Assembly	Analytic-autonomizing type	When encountering difficulties he first responds with analytical, logical and restraint approach, followed by an unyielding, powerful attack, based on logic and planning. If these approaches do not work, he gives up for the sake of harmony, but he chooses that as the last resort.	He functions in a team, although individual work suits him more.
8 Planning	Cautious-supporting type	He strives most for harmony and readiness for cooperation. If he does not achieve this, he tries to withdraw and save what he can. If even this is not successful, he quarrels, most probably very explosively.	He functions in a team, although individual work suits him more.

Table 7. Personal value systems of team members

Analysis of results reveals that in this team three members have very strong feeling for team work, two of them have good feeling for team work, while three members function in a team although individual work suits them more.

The prevailing value system of team members is cautious-supporting, bordering on analytic-autonomizing and flexible value system. No team member has either altruistic-nurturing or assertive-directing value system.

3.3 Analysis of representation of team roles

Nine team roles have to be represented in a team for successful team-work (Belbin, 2003): developer, searcher for resources, coordinator, creator, evaluator, co-worker, operator, finisher and expert.

In order to find out whether all nine team roles are represented in the proposed team for project-driven concurrent development of the car pedal component, team members got Belbin test questionnaire.

Answers of team members to four self-evaluation questionnaires and evaluation of team co-workers were used as input data for INTERPLACE software. This software was used to find out which are the natural roles of each team member, which roles can he adapt to and which roles he should avoid (Table 8).

Team member	Team roles								
	DE	SR	CO	CR	EV	CW	OP	FI	EX
1 Control	x		x				x		
2 Development	x	x			x				
3 Sales				x			x		x
4 Project manager			x	x			x		
5 Supply			x		x				x
6 Process planning	x						x		x
7 Assembly					x		x		x
8 Planning						x	x	x	

Legend:

DE – developer	CW – co-worker
SR – searcher for resources	OP - operator
CO – coordinator	FI - finisher
CR – creator	EX - expert
EV – evaluator	x – natural role

Table 8. An overview of the natural roles of eight team members:

It can be seen from Table 8 that all nine team roles are represented in the team. This indicates the capability of the team for efficient team-work in the project of concurrent development of pedal component.
After the presentation of analyses results, the company management decided that the selected and tested eight-member team would start working on the project for concurrent development of pedal component.

4. Conclusion

Global market requires short product development times and low costs, and this forces companies to a transition from sequential to concurrent product and processes development.
Prerequisite for a transition from sequential to project-driven concurrent product development is a successful team-work and strategic management (parallelness, standardisation and integration). In our paper we have therefore focused on checking the efficiency of team-work and integration of strategic management into product development processes.
Results of the analysis of team-work capability, team work motivation, susceptibility of the team to dysfunctions, personal value systems for team-work and representation of team roles led us to the conclusion that we propose to the company management to include the treated and tested team in the project of concurrent development of the pedal component.
Further we analysed changes in a company, which were required by a transition from sequential to concurrent product development, with respect to:

- organisational concept of the company,
- organisation of processes,
- organisation of work,
- IT,
- preparation of product documentation.

5. References

Belbin, R., Meredith (2003). *Team roles at work*, Elsevier, ISBN 0-7506-2675-5, Amsterdam.

Bullinger, H.,J. & Warnecke, H.,J. (1996). *Neue Organisationsformen in Unternehmen*, Springer-Verlag, ISBN 3-540-60263-1, Berlin Heidelberg.

Duhovnik, J.; Starbek, M.; Dwivedi, S.,N. & Prasad, B. (2001). Development of New Products in Small Companies, *Concurrent engineering: Research and Applications,* Vol. 9, No. 3, (september 2001), pp 191-210, ISSN 1063-293x.

Ellis, A.;Wallis, P. & Washburn, S. (2006). *Charting your course for effective communication: SDI and communication*, Personal Strengths Publishing, ISBN 1-932627-04-9, Carlsbad.

Eversheim, W.; Bochtler, W. & Laufenberg, L. (1995). *Simultaneous Engineering*, Springer-Verlag, ISBN 3-540-57882x, Berlin Heidelberg.

Kušar, J.; Bradeško, L.; Duhovnik, J. & Starbek, M. (2008). Project management of product development. *Journal of Mechanical Engineering*, Vol. 54, No. 9, (september 2008), pp. 588-606, ISSN 0039-2480.

Lencioni, P. (2002). *The Five Dysfunctions of a Team*, Jossey-Bass, ISBN 0-7879-6075-6, San Francisco.

Prasad, B. (1996). *Concurrent Engineering Fundamentals, Volume I, Integrated Product and Process Organization*, Prentice Hall PTR, ISBN ,0-13-147463-4, New Jersey

7

Forecasting of Production Order Lead Time in Sme's

Tomaž Berlec and Marko Starbek
Laboratory for Manufacturing Systems and Production Process Planning,
Faculty of Mechanical Engineering, University of Ljubljana
Slovenia

1. Introduction

More and more companies on the global market are today capable of manufacturing individual or small-series orders at comparable prices and quality. The main difference between these companies is the expected production order development-time and the observance of delivery deadlines.
Before making a bid, the sales department must establish what operations will have to be carried out for a particular order, the time needed for performing these operations, and what delivery time is required.
Currently, operation-time data are usually obtained from experienced company employees (however, a problem arises if they leave the company, because SMEs do not usually have systems for knowledge capture—the "knowledge" with which they are dealing is more "oral tradition", knowledge obtained by experience), while the customer specifies the delivery time. However, estimates on lead times, and thus delivery times, based on personal experience can be rather misleading. Bids may consequently be based on wrong decisions, or even worse; because of an incorrectly specified delivery time the company may not receive the order, because the delivery time (if not specified by the customer) may be too long and therefore uncompetitive in comparison with other bids. Another type of a problem arises if the specified delivery time is too short and cannot be met.
The development of information and communication technologies (ICT) makes it easier for a company to improve and maintain its competitive advantages on the market (Leem C.S., Suh J.W., 2005), because it is very easy to access the data. A company striving to be competitive on the global market needs a suitable enterprise resource planning (ERP) system. There are several ERP systems available (Scherer E., 2005) and each company must select the optimal system for its needs (Starbek M., Grum J., 2000).
This chapter presents how the data stored in the ERP system can be used for the calculation of lead times of operational and assembly orders and indirectly, for forecasting production order lead times, depending on the confidence interval.
Naturally, if the company does not have an ERP system, it is possible to manually obtain the data required for forecasting lead times. The disadvantage of such a method is that the manual procedure is rather time-consuming in order to build an applicable database.
The chapter presents a review of the literature and some achievements and guidelines related to lead times of orders and delivery times. A procedure for forecasting production

order lead times is presented and described, as well as the results of the application of this procedure in a tool shop.
The result of the proposed procedure for forecasting production order lead times is an empirical distribution of possible lead times for a production order. On the basis of this distribution, it is possible to forecast the probable lead time of a production order as a function of the confidence interval.
Using the proposed procedure, the sales department can make a delivery time forecast for the customer of the planned production order.
Conclusions, findings and guidelines for further activities are presented at the end of the chapter.

2. Literature review

Considerable research has been done on the possibilities of determining production lead times. In 1979, Weeks (Weeks, 1979) researched the impact of forecasted lead times based on various statistical measurements in individual production of variable volume and structure. He tested the following three hypotheses:
Delivery time rules based on estimates of individual job lead-time conditions have a better effect on workshop congestion than widely reported total work content rules when employed with delivery time oriented dispatching rules.
Delivery time oriented dispatching rule investigated performed better than the shortest-imminent-processing-time dispatching rule in terms of meeting delivery times.
Delivery time performance tends to worsen as workshop structure becomes more elaborated and complex.
Weeks concluded that this was just the beginnings of research on forecasting delivery lead times with one of several possible statistical tools and that there were many unanswered questions and much research would have to be done in this field.
(Vig & Dooley, 1991) used two new dynamic rules for defining delivery time in existing delivery-time forecast models. They discovered that data on orders that had been completed recently could be very useful for forecasts in the future. Their study confirmed the conclusions of other research: the characteristics of a particular order and the type of production are very important for the forecast of lead time.
(Enns, 1994) stated that short lead times and high supply reliability were required for job shop customers.
Lawrence S.R. (1994) presented a methodology for negotiating due dates between the customer and the producer in a complex production environment.
Nyhuis P., Vogel M. (2006) presented a methodology for tracking and accurate logistic control of a one-piece material-flow process.
Denkena B., Lorenzen L.-E., Batino A. (2006) studied the possibilities of increasing production flexibility and efficiency and proposed a new production-planning model: integration of centralized labour and decentralized decision-making.
Several studies (Buzacott J.A. & Mandelbaum M., 1985; Chen Y.J. et al., 2005; Wang Z. et al., 2004; Krause F.L. & Altmann C., 1991) have shown that the flexibility of enterprise resource planning can be improved only if alternative technology solutions are used during repeated planning of manufacturing operations. However, the aforementioned studies do not deal with the basic question of how to obtain quality input data for successful production planning.

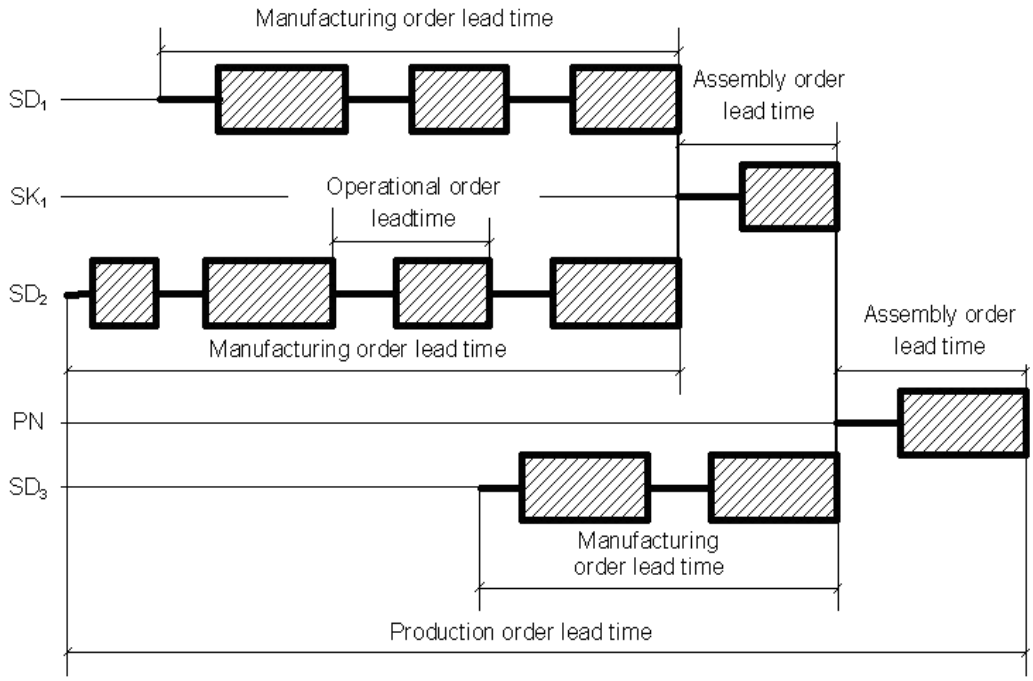
Because of the ever more dynamic market, Wiendahl H.-P., Dammann M., 2006, presented the concept of a method for measuring dynamic influences and tools that can help companies to choose the right response to these dynamic influences. He used Begemann's (Begemann C., 2006) approach for capacity control, targeted at completion time, and Lödding's (Lödding H., 2005) model of production control. However, according to Wiendahl, the whole concept is still in its initial phase and needs detailed overview and research.
In our research, we did not find any lead time forecasting approach as described in this paper, so we assume this is a new approach that uses known theory on lead times and adds a new method for forecasting production order lead times.
Tatsiopoulos I.P.& Kingsman B.G., 1983, presented a comparison of two alternative approaches for determining planned values for manufacturing lead times for use in production planning and control systems. One approach was to treat manufacturing lead times as probabilistic and the second was to emphasise the control of manufacturing lead times. The conclusion of their paper was that new tools would have to be developed for planning, and theories for determining lead times would have to be improved.
Kingsman B.G., et al., 1989, described a developed methodology for controlling manufacturing lead times in make-to-order companies.
Kingsman B.G. (2000) presented modelling of input-output workload control for dynamic capacity planning in production planning systems, and ended with the conclusion that the arrival of orders in produce-to-order companies cannot be forecast in advance, and that managing lead times is a better approach than using forecast lead times. These conclusions encouraged us to try to find a better way of forecasting production order lead times.
Ooijen & Bertrand, 2001, wrote that, from the economic point of view, it is necessary to process orders within the deadline and, at the same time, it is necessary to take into account the acceptance of the delivery date from the customer's point of view and consider the reality of the deadline when making a bid.
Over the past decade, a lot of advanced methods and generic algorithms for scheduling production processes (Lestan, et al., 2009; Tasic et al., 2007; Kušar et al., 2004) and scheduling systems have been developed in industry and academia, but there are still unsolved general problems.
Studies on common cycle lot-size scheduling for a multi-product and multi-stage arborescent flow-shop environment was done (Ashjari & Fatemi Ghom, 2001; Fatemi Ghomi & Torabi, 2001) and solution methods to determine simultaneous production cycle time and production schedule were given.
Öztürk et al., 2006, found that it is not enough to forecast short delivery times—the forecasted delivery times has to be accurate. The main problem is that most of the lead time consists of queue and transport, while a relatively small part of it consists of the actual processing—which is why it is so difficult to forecast lead times. They tried to forecast lead times by using data mining; they used a regression tree approach and linear regression for forecast.
In our research, we did not find any lead time forecast approach as described in this chapter, so we assume that this is a new approach that uses known theory on lead times and adds a new method for forecasting production order lead times.

3. How to FORECAST a production order lead time

When dealing generally with "an order", it is necessary to distinguish between (Wiendahl H.P., 1995):

- operational order,
- manufacturing order,
- assembly order,
- production order.

The types of orders given above and their corresponding lead times are shown in Figure 1.



Legend:

PN – production order
SK_x – x-th component
SD_x – x-th part

Fig. 1. Types of orders and their corresponding lead times (Wiendahl H.P., 1995)

When designing a procedure for forecasting production order lead times, it will be assumed that the company wishing to forecast the lead times of orders uses an ERP/PPC system, the database of which contains data on past operational and assembly orders.
The ERP/PPC system will be the basis for forecasting production order lead times.

4. Method for FORECASTING production order lead times

An overview of known procedures in the literature for determining realistic lead times of operations (Wiendahl H.P., 1995; Nyhuis P., Wiendahl H.P., 1999) and the experience obtained during many tests of practical implementation of these procedures, led us to the conclusion that it would be possible to forecast lead times of planned orders on the basis of actual operational and assembly order lead times achieved in the past. These forecasts (on the basis of ERP-system data or on the basis of manually acquired past data) are accurate enough for individual production because manufacturing processes on individual machines

are taken into account. By using these data it is thus possible to forecast lead times even for fairly complex products with several machining operations and individual order features. Based on our research we concluded that the procedure for forecasting lead times for future production orders should consist of the steps shown in Figure 2.

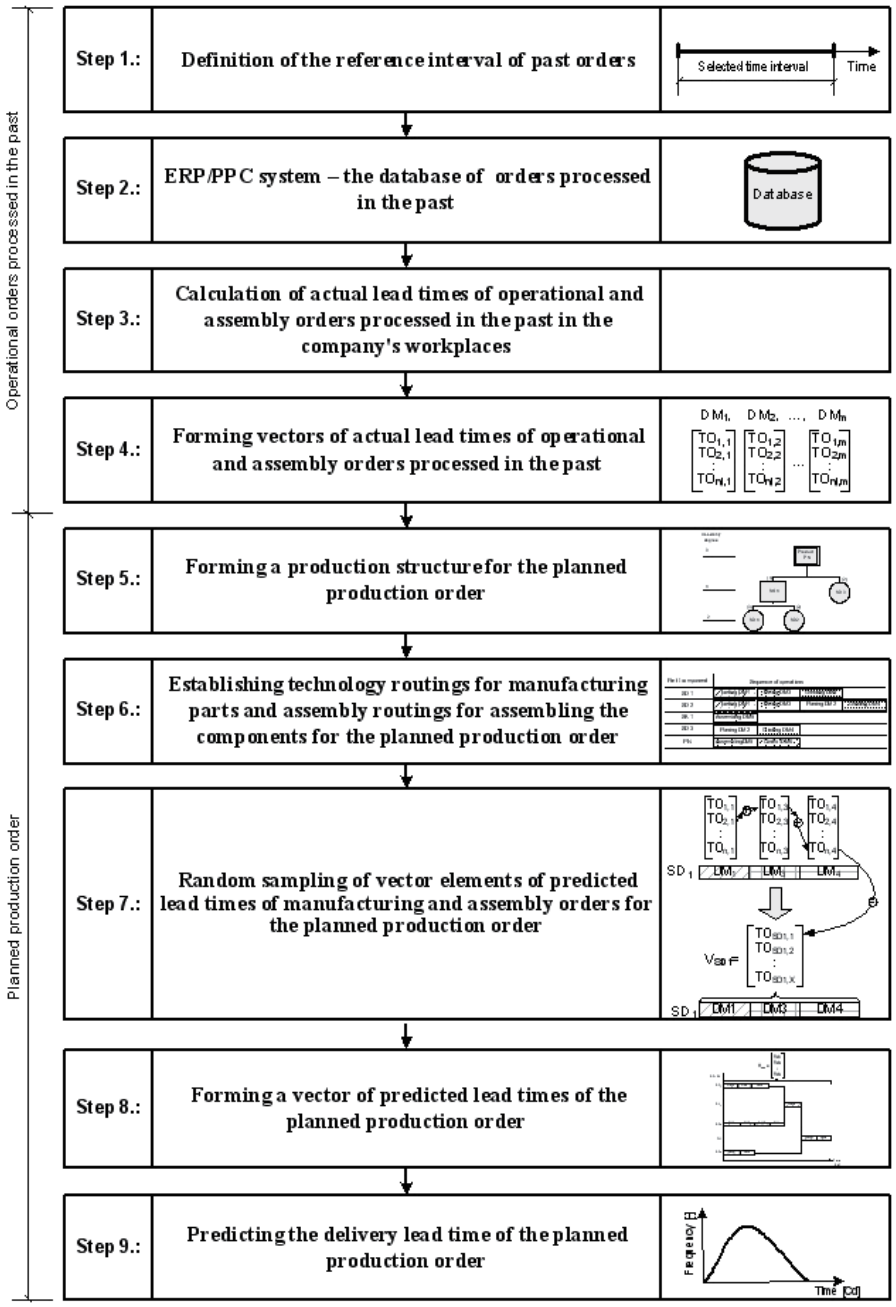


Fig. 2. Procedure for predicting planned production order lead times

Step 1. Definition of the reference interval of past orders

At the beginning of lead time forecasting it is necessary to define the interval for data acquisition of past operational and assembly orders. This interval can be a month, a quarter, a year or several years.

Step 2. ERP/PPC system – the database of orders processed in the past

As mentioned, a company wishing to forecast lead times must have an ERP/PPC system as a basis for all further steps, because this is the database of orders processed in the past.
The ERP/PPC system should provide data on (Figure 3):

- operational or assembly order codes,
- type and sequence of operations in manufacturing and assembly orders,
- IDs of workplaces at which operational or assembly orders have been processed,
- actual execution times of operational or assembly orders,
- date of completing a particular operational or assembly order in the previous workplace,
- date of completing a particular operational or assembly order in the observed workplace.

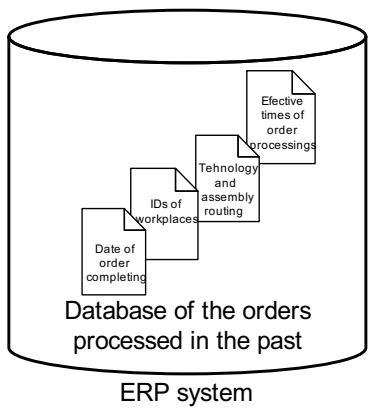


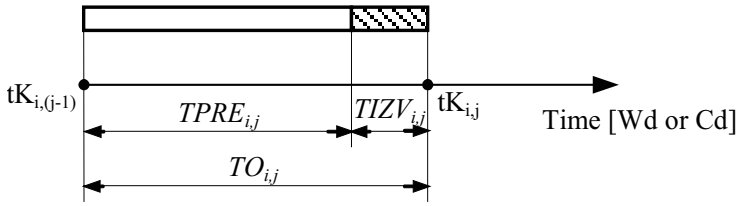
Fig. 3. ERP/PPC system database

Step 3. Calculation of actual lead times of operational and assembly orders processed in the past in the company's workplaces

The lead time of the *i*-th operational order N_i $(1 \le i \le n)$, processed in the *j*-th workplace DM_j $(1 \le j \le m)$ is defined as the interval calculated from the time when the *i*-th operational order was completed in the previous, i.e. *(j-1)*-th workplace, until the time when the *i*-th operational order is completed in the observed, i.e. *j*-th workplace (Wiendahl H.P., 1995), as presented in Figure 4.
The lead time of an operational or assembly order is therefore:

$$TO_{i,j} = tK_{i,j} - tK_{i,(j-1)} \tag{1}$$

$TO_{i,j}$ – lead time of the i-th operational order in the j-th workplace
$tK_{i,j}$ – completion time of the i-th operational order in the j-th workplace
$tK_{i,(j-1)}$ – completion time of the i-th operational order in the previous (j-1)-th workplace



Legend:

TPRE$_{ij}$ – crossing time of the *i*-th operational order in the *j*-th workplace
TIZV$_{ij}$ – execution time of the *i*-th operational order in the *j*-th workplace
TO$_{i,j}$ – lead time of the i-th operational order in the *j*-th workplace
tK$_{i,j}$ – completion time of the *i*-th operational order in the *j*-th workplace
tK$_{i,(j-1)}$ – completion time of the *i*-th operational order in the previous *(j-1)*-th workplace
Wd – work day
Cd – calendar day

Fig. 4. Lead time of operational order (Wiendahl H.P., 1995)

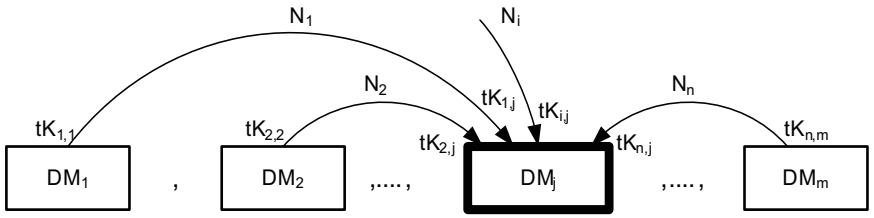


Fig. 5. Flow of operational orders through DM_j workplace

On the basis of the ERP/PPC system output data, it is possible to calculate (for any *j*-th workplace DM_j) the actual lead times of previously processed operational orders, i.e. orders that have been processed in the *j*-th workplace in the observed time interval (Figure 5).
The actual lead times of operational orders processed in the *j*-th workplace in the selected time interval are therefore:

$$TO_{1,j} = tK_{1,j} - tK_{1,(j-1)}$$

$$TO_{2,j} = tK_{2,j} - tK_{2,(j-1)}$$

$$\vdots$$

$$TO_{i,j} = tK_{i,j} - tK_{i,(j-1)}$$

$$\vdots$$

$$TO_{nj,j} = tK_{nj,j} - tK_{nj,(j-1)} \qquad (2)$$

Step 4. Forming vectors of actual lead times of operational and assembly orders processed in the past

It is necessary to form vectors of actual lead times of orders processed in the past in the company's workplaces (Table 1).

	SELECTED INTERVAL from … to … [Wd]					
Workplace	DM 1	DM 2	…	DM j	…	DM m
Vectors of actual lead times of orders	$\begin{bmatrix} TO_{1,1} \\ TO_{2,1} \\ . \\ . \\ . \\ TO_{n_j,1} \end{bmatrix}$	$\begin{bmatrix} TO_{1,2} \\ TO_{2,2} \\ . \\ . \\ . \\ TO_{n_j,2} \end{bmatrix}$	…	$\begin{bmatrix} TO_{1,j} \\ TO_{2,j} \\ . \\ . \\ . \\ TO_{n_j,j} \end{bmatrix}$	…	$\begin{bmatrix} TO_{1,m} \\ TO_{2,m} \\ . \\ . \\ . \\ TO_{n_j,m} \end{bmatrix}$

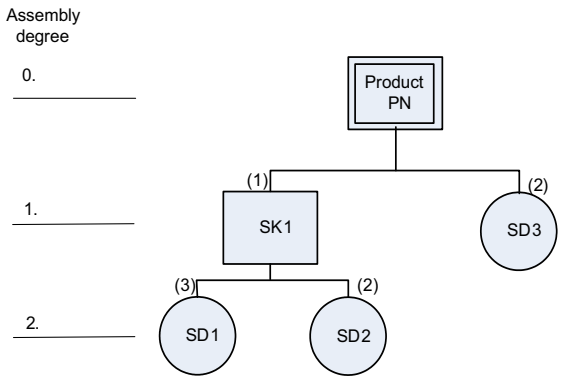
Legend:
$TO_{i,j}$ – lead time of the *i*-th operational order in the *j*-th workplace
DM_j – *j*-th workplace
Wd – work day

Table 1. Vectors of actual lead times of operational and assembly orders processed in the past

Vectors of actual lead times of orders processed in the past will be the basic data for forecasting lead times of the planned new production orders.

Step 5. Forming a production structure for the planned production order

It is necessary to make a graphic presentation of the production order structure for the planned production order (Figure 6).



Legend: – production order
– component
– part
(x) – the number of times a part or component was built into a higher-level component

Fig. 6. Production structure of the planned production order

Step 6. Establishing technology routings for manufacturing parts and assembly routings for assembling the components for the planned production order

Figure 7 gives an overview of technology and assembly routings for manufacturing parts and assembling components of the planned production order.

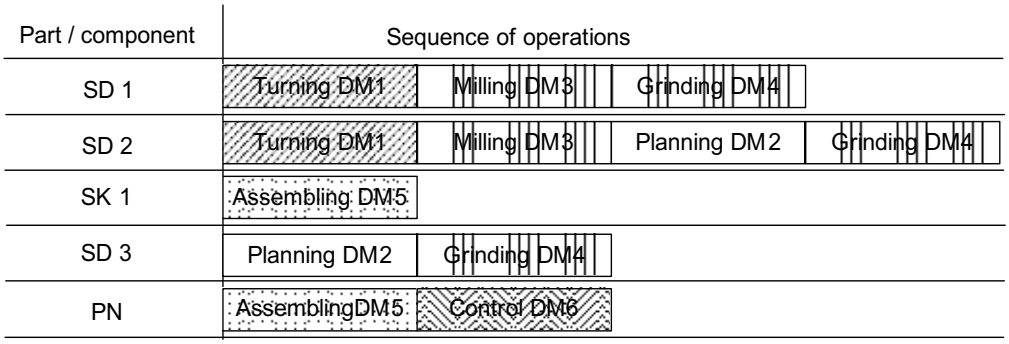


Fig. 7. Technology and assembly routings for manufacturing parts and assembly of components of the PN production order

Step 7. Random sampling of vector elements of forecast lead times of manufacturing and assembly orders for the planned production order

On the basis of the production structure (defined in step 5) for the planned production order, and on the basis of technology and assembly routings for manufacturing parts and assembling components (defined in step 6), software (SPSS, Matlab, etc.) can be used to form vectors of forecast lead times of manufacturing and assembly orders for the planned production order (Figure 8).

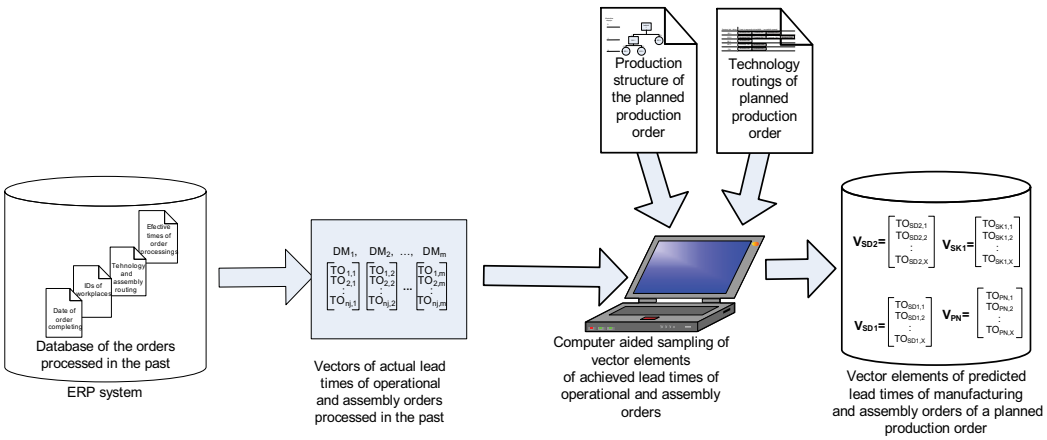


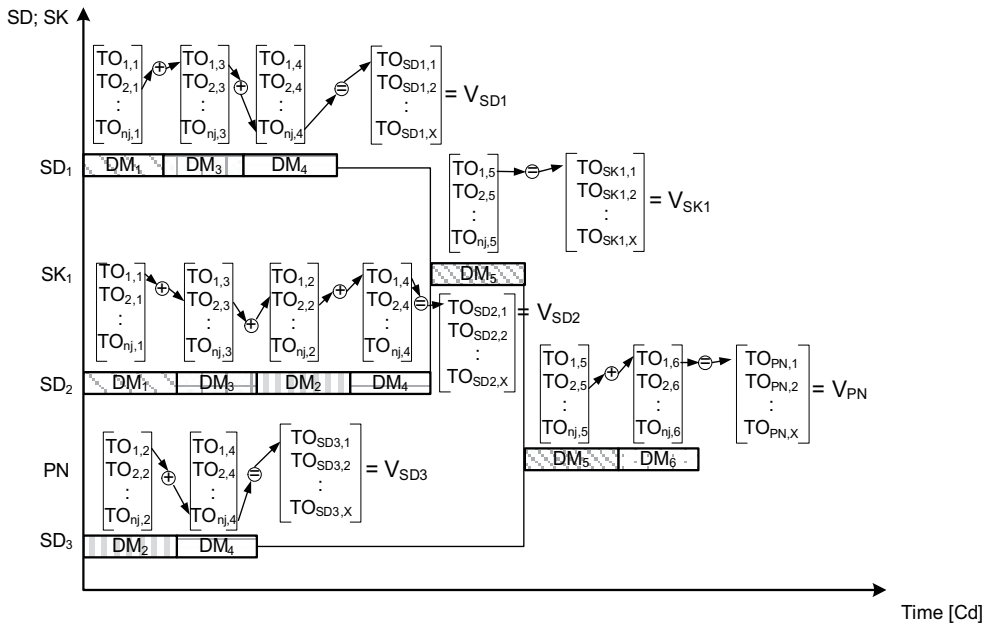
Fig. 8. Principle of vector element sampling of predicted lead times for manufacturing and assembly orders of a planned production order

Figure 8 shows the principle of computer-aided sampling of vector elements of forecast lead times for manufacturing and assembly orders for the planned production order.
The results of sampling are vectors of random lead times of parts and components. The number of vector elements of forecast lead times of manufacturing and assembly orders for

the planned production order depends on the number of random samplings performed on a random selection of lead times for manufacturing and assembly orders for the planned production order. Tests have shown that lead times with a small number of random samplings (500 samplings) differ considerably from lead times obtained with a higher number of random samplings (5000 samplings). A large further increase in sampling number (50000 samplings) does not significantly change the results, it merely increases the computing time.

Tests will therefore be required to define the number of random sampling in order to ensure a stable process.

Figure 9 is a Gantt chart presentation of the principle of random sampling of vector elements of forecast lead times of manufacturing and assembly orders for a planned production order.



Legend: ⊕ – the summing symbol

$TO_{SD1,1}$ – lead time of the SD_1 part, obtained after the first sampling

$TO_{SK1,1}$ – lead time of the SK_1 component, obtained after the first sampling

$TO_{PN1,1}$ – lead time of the PN product, obtained after the first sampling

V_{SD1} – vector of lead times of the SD_1 part, obtained after x samplings

Fig. 9. Gantt chart of random sampling of vector elements of predicted lead times for manufacturing and assembly orders of the planned production order

Step 8. Forming a vector of forecast lead times of the planned production order

In order to define the vector elements of forecast lead times for the planned production order, the Gantt chart of a production order (Figure 9) must be transformed into an activity network diagram for the production order and entered into the lead times (found during sampling in step 7) of parts and components for the planned production order (Figure 10).

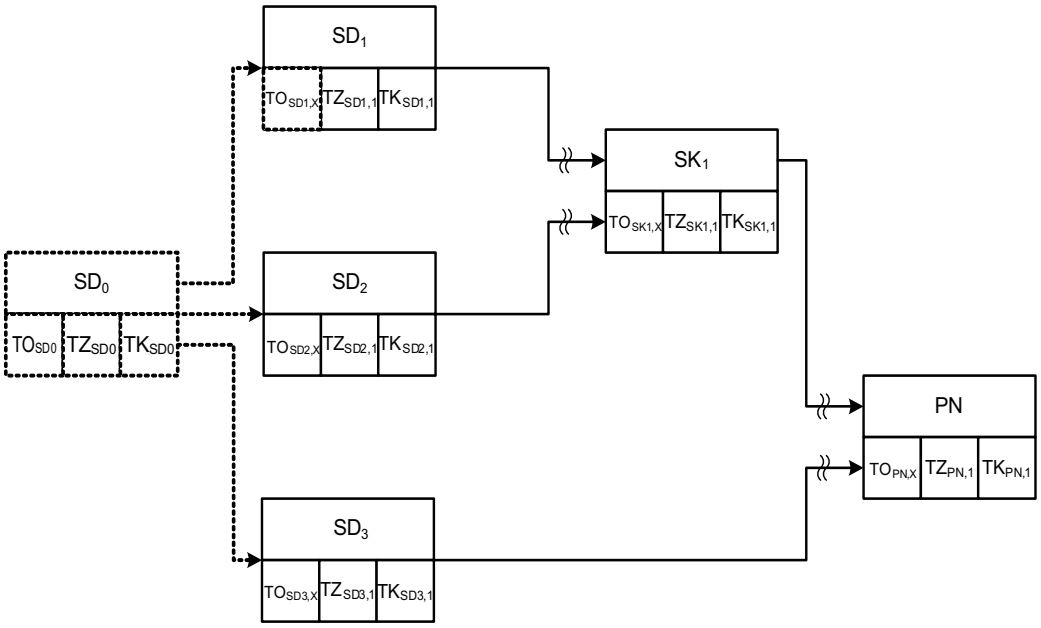


Fig. 10. Activity network diagram of the planned production order

Initial data for the activity network diagram of the production order:

- date of starting the processing of the virtual manufacturing/assembly order SD_0

$$TZ_{SD_0} = 0 \tag{3}$$

- vector of virtual manufacturing/assembly order V_{SD0}:

$$V_{SD_0} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix} \tag{4}$$

vectors of the expected lead times of manufacturing/assembly orders for the planned production order:

$$V_{SD_1}, V_{SD_2}, \dots, V_{SK_1}, V_{PN}$$

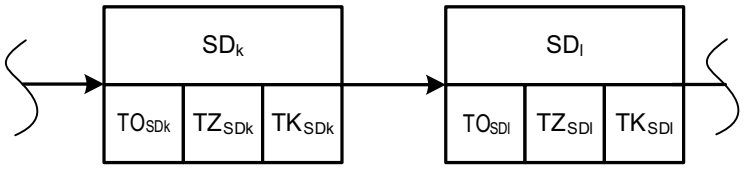
For the virtual manufacturing/assembly order SD_0, which has no predecessors in the activity network diagram, it is assumed that the date of starting the order processing is

$$TZ_{SD_0} = 0 \tag{5}$$

The date of completing the order processing is

$$TK_{SD_0} = TZ_{SD_0} + TO_{SD_0} = 0 + 0 = 0\,. \tag{6}$$

For other manufacturing or assembly orders which have one or more predecessors (Figure 11):



SD_k – *k*-th part
TO_{SDk} – lead time of the SD_k part
TZ_{SDk} – date o f starting the processing of the manufacturing or assembly order on the SD_k part
TK_{SDk} – date of completing the processing of the manufacturing or assembly order on the SD_k part
SD_l – *l*-th part
TO_{SDl} – lead time of the SD_l part
TZ_{SDl} – date of starting the processing of the manufacturing or assembly order on the SD_l part
TK_{SDl} – date of completing the processing of the manufacturing or assembly order on the SD_l part

Fig. 11. Basic elements of activity network diagram

The date of starting the processing of the *l*-th manufacturing or assembly order:

$$TZ_{SD_l} = \max_{k \in PR} \left\{ TZ_{SD_k} + TO_{SD_{k,X}} \right\} \tag{7}$$

PR – predecessors of the observed order l
The date of completing the processing of the l-th manufacturing or assembly order:

$$TK_{SD_l} = TZ_{SD_l} + TO_{SD_{l,X}} \tag{8}$$

The date of completing the last assembly order in the activity network diagram is equivalent to the expected lead time of the planned production order TO

$$TK_{PN} = TO\,. \tag{9}$$

Figure 10 shows the calculation for one vector element of the expected lead time of the planned production order. Such a calculation has to be repeated for a selected number of iterations of randomly sampled values from vectors of an individual part or component of a production order, which finally leads to the vector of the forecast lead times of the planned production order and corresponding distribution function of the order lead time.

Step 9: Forecasting the delivery lead time of the planned production order

The result of step 8 of the procedure for forecasting the production-order lead time is the vector of forecast lead times of the planned production order and the corresponding order lead time distribution function.
In real life, however, an exact deadline for product delivery to the customer is required.
The most probable delivery lead time for the planned production order can be estimated by using the median, which means that there is a 50% probability that the actual delivery time will be shorter, and 50% probability that it will be longer than forecast.

A 50% probability is not acceptable in practice, so it is necessary to extend the confidence interval and thus also the forecast production order lead time.
In engineering, a 95% confidence interval is usually used, which means that there is a 95% probability that the production order will be delivered within the forecast lead time.
The maximum order-delivery lead time that can be guaranteed to the customer with a 95% probability, therefore corresponds to the 95th percentile of the empirical distribution of the forecast lead time of the production order (Rice J. A., 1995; The MathWorks, Inc., 2002), as shown in Figure 12.

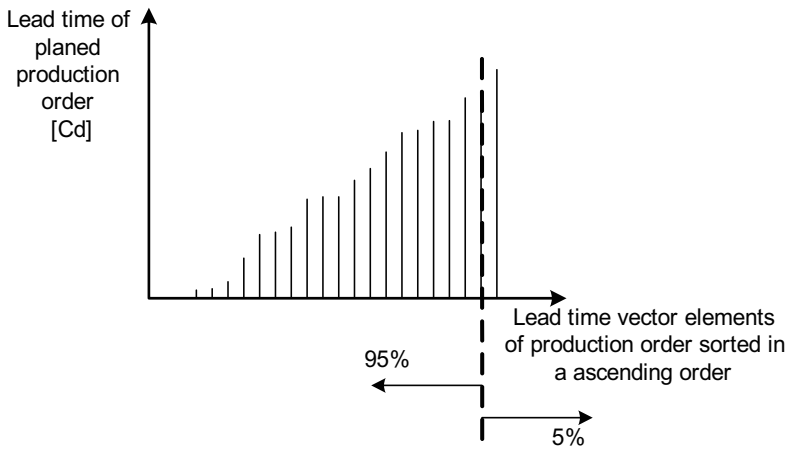


Fig. 12. An example of the 95th percentile

In order to obtain the *P*-th percentile of lead time vector elements (*X*), sorted in an ascending order, it is necessary to calculate the percentile rank *R* (Ferligoj A., 1995):

$$R = \frac{X}{100} \cdot P + \frac{1}{2} \tag{10}$$

R – percentile rank—sequence number of an element in the lead time vector sorted in an ascending order
P – percentile
This value is rounded to the nearest integer and the value from the X set which corresponds to this rank is then selected.
Naturally, the choice of the percentile may depend on the importance of the order and the customer; the more important the customer, or the more important the order, the stronger is the interest of the company in obtaining a particular order; so the company will be ready to accept a higher risk.
After all nine steps of the procedure have been completed, a good forecasting of the production order lead time can be obtained, which is then sent to the customer.

5. Testing the procedure for FORECASTING production order lead time

The procedure for predicting production order lead times was tested in a tool shop company from Slovenia. The company produces tools for transforming and cutting, tools for injection moulding of thermoplastic and duroplastic materials, jet and press machines for

duroplastic materials, press machines for ceramic materials, and automated assembly appliances. The model for forecasting lead times was tested in the tool shop for manufacturing tools, but not for manufacturing devices. A separate database for manufacturing devices would have to be made, because orders for tools are completely different from orders for devices.

The tool shop's speciality is designing and manufacturing high-quality tools for injection moulding of thermoplastic and duroplastic materials. On the basis of its long experience in making tools for its parent company, the tool shop started producing tools and appliances for external customers in the following fields:

- automotive industry,
- household appliances,
- medical technology,
- electrical engineering and electronics,
- illumination.

The tool shop uses the Largo ERP system, developed by the Perftech Company from Bled, Slovenia (Largo, 2007). Due to their production method (tools are made for known customers and each tool is unique) it is very difficult to precisely forecast the duration of tool production, yet this information is essential for making bids and winning orders. In the past, delivery times were guessed or were estimated by experienced company employees. Several times the specified delivery times were too short, which resulted in penalties, sometimes even in cancellation of further cooperation with a particular business partner; and goodwill was also affected. In all industries (but especially the automotive), it is very important that the agreed deadlines are met, because SMEs are usually sub-suppliers or suppliers in a long supply-chain for a large corporation and if one delivery is late, the whole supply chain may be late.

The company management therefore decided to test the suitability of the proposed procedure for forecasting lead times of production orders in a case study of determining the lead time of a production order for a "tool for a linking element of an oil vent # 708145". The final product manufactured with this tool is shown in Figure 13.

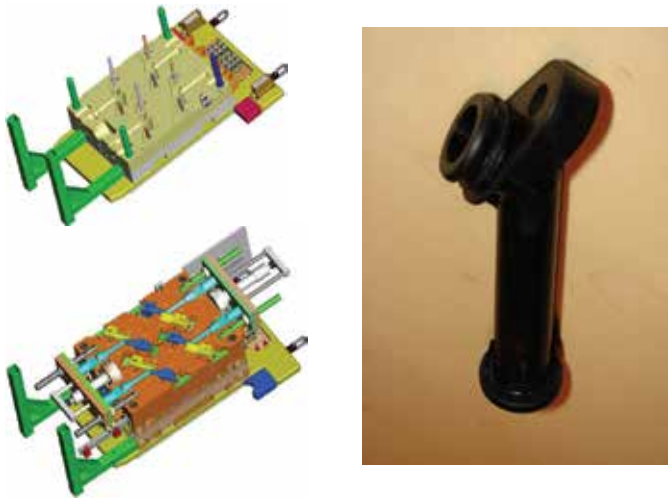


Fig. 13. Picture of tool # 708145 and the linking element of an oil vent made with this tool

Steps of the procedure for forecasting the production order lead time for the "tool for a linking element of an oil vent # 708145":

Step 1. Definition of the reference interval of past orders

In agreement with the tool shop management, it was decided that data from 12 December 2002 to 22 August 2005 would be used for determining the actual lead times of operational orders in the past.

Step 2. ERP/PPC system – the database of orders processed in the past

The data from the Largo ERP system database were first transformed to MS Excel format.
The following data were used from the ERP database: order number, arrival date, departure date, manufacturing time, and technology and assembly routings.
The Largo ERP system uses calendar dates and does not take into account the company's workday calendar.
During the time defined in step 1, 22,850 manufacturing orders were processed, with 57,951 operational orders in 35 workplaces (Table 2).

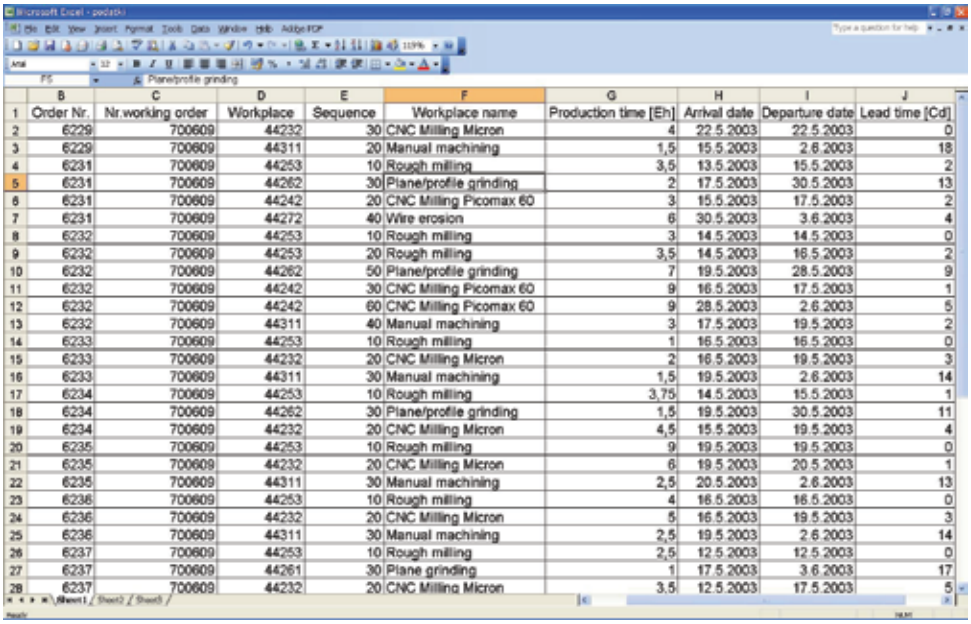
Workplace number	Workplace	Number of orders processed in 3 years
44000	Cooperation – service	21
44141	Design of devices	151
44142	Machine electronics	130
44143	Design of tools	2288
44211	Slitting	1420
44221	Turning	3706
44222	CNC turning	1052
44231	CNC programming	371
44232	CNC Micron milling	2660
44241	CNC programming	668
44242	CNC Picomax 60 milling	4153
44291	Heat treatment	5172
44311	Manual machining	4288
44312	Assembly of tools	812
44313	Assembly of machines and devices	197
44321	Sampling	2
44331	Measurement	885
44332	DEA Omicron measurement	273
	3-year production:	57951

Table 2. Number of operational orders processed in the tool shop workplaces

It can be seen from Table 2 that a widely varying number of operational orders were processed in workplaces during the observed time (minimum 2 orders in workplace 44,321 and maximum 7307 orders in workplace 44,253).

Step 3. Calculation of actual lead times of operational and assembly orders processed in the past in the company's workplaces

Actual lead times of individual operational orders were calculated from the data obtained in steps 1 and 2. The calculation was done in MS Excel, using equation 1.
Figure 14 shows part of the calculation of actual lead times of operational orders in an Excel table.



	B	C	D	E	F	G	H	I	J
1	Order Nr.	Nr.working order	Workplace	Sequence	Workplace name	Production time [Eh]	Arrival date	Departure date	Lead time [Cd]
2	6229	700609	44232	30	CNC Milling Micron	4	22.5.2003	22.5.2003	0
3	6229	700609	44311	20	Manual machining	1,5	15.5.2003	2.6.2003	18
4	6231	700609	44253	10	Rough milling	3,5	13.5.2003	15.5.2003	2
5	6231	700609	44262	30	Plane/profile grinding	2	17.5.2003	30.5.2003	13
6	6231	700609	44242	20	CNC Milling Picomax 60	3	15.5.2003	17.5.2003	2
7	6231	700609	44272	40	Wire erosion	6	30.5.2003	3.6.2003	4
8	6232	700609	44253	10	Rough milling	3	14.5.2003	14.5.2003	0
9	6232	700609	44253	20	Rough milling	3,5	14.5.2003	16.5.2003	2
10	6232	700609	44262	50	Plane/profile grinding	7	19.5.2003	28.5.2003	9
11	6232	700609	44242	30	CNC Milling Picomax 60	9	16.5.2003	17.5.2003	1
12	6232	700609	44242	60	CNC Milling Picomax 60	9	28.5.2003	2.6.2003	5
13	6232	700609	44311	40	Manual machining	3	17.5.2003	19.5.2003	2
14	6233	700609	44253	10	Rough milling	1	16.5.2003	16.5.2003	0
15	6233	700609	44232	20	CNC Milling Micron	2	16.5.2003	19.5.2003	3
16	6233	700609	44311	30	Manual machining	1,5	19.5.2003	2.6.2003	14
17	6234	700609	44253	10	Rough milling	3,75	14.5.2003	15.5.2003	1
18	6234	700609	44262	30	Plane/profile grinding	1,5	19.5.2003	30.5.2003	11
19	6234	700609	44232	20	CNC Milling Micron	4,5	15.5.2003	19.5.2003	4
20	6235	700609	44253	10	Rough milling	9	19.5.2003	19.5.2003	0
21	6235	700609	44232	20	CNC Milling Micron	6	19.5.2003	20.5.2003	1
22	6235	700609	44311	30	Manual machining	2,5	20.5.2003	2.6.2003	13
23	6236	700609	44253	10	Rough milling	4	16.5.2003	16.5.2003	0
24	6236	700609	44232	20	CNC Milling Micron	5	16.5.2003	19.5.2003	3
25	6236	700609	44311	30	Manual machining	2,5	19.5.2003	2.6.2003	14
26	6237	700609	44253	10	Rough milling	2,5	12.5.2003	12.5.2003	0
27	6237	700609	44261	30	Plane grinding	1	17.5.2003	3.6.2003	17
28	6237	700609	44232	20	CNC Milling Micron	3,5	12.5.2003	17.5.2003	5

Fig. 14. Calculation of actual lead times of operational orders processed from 12 December 2002 to 22 August 2005

The results showed that the majority of actual lead times are shorter than or equal to 1 calendar day (Cd). Some extreme cases, e.g. 464 Cd, are exceptions to the rule.

Step 4. Forming vectors of actual lead times of operational and assembly orders processed in the past

The results obtained in step 3 were transformed into vectors of actual lead times; part of the data is shown in Table 3.

	SELECTED INTERVAL from 12 Dec 2002 till 22 Aug 2005					
Workplace	44000	44141	…	44253	…	44332
Vectors of actual lead times of orders	$\begin{bmatrix} 0 \\ 5 \\ . \\ . \\ . \\ 2 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 66 \\ . \\ . \\ . \\ 9 \end{bmatrix}$	…	$\begin{bmatrix} 7 \\ 4 \\ . \\ . \\ . \\ 1 \end{bmatrix}$	…	$\begin{bmatrix} 2 \\ 10 \\ . \\ . \\ . \\ 0 \end{bmatrix}$
Number of vector elements	21	151		7307		273

Table 3. Vectors of actual lead times of operational and assembly orders processed in the past

Step 5. Forming a production structure for the planned production order – tool # 708145

In this step, the known production/assembly structure of tool # 708145 was used (Figure 15).

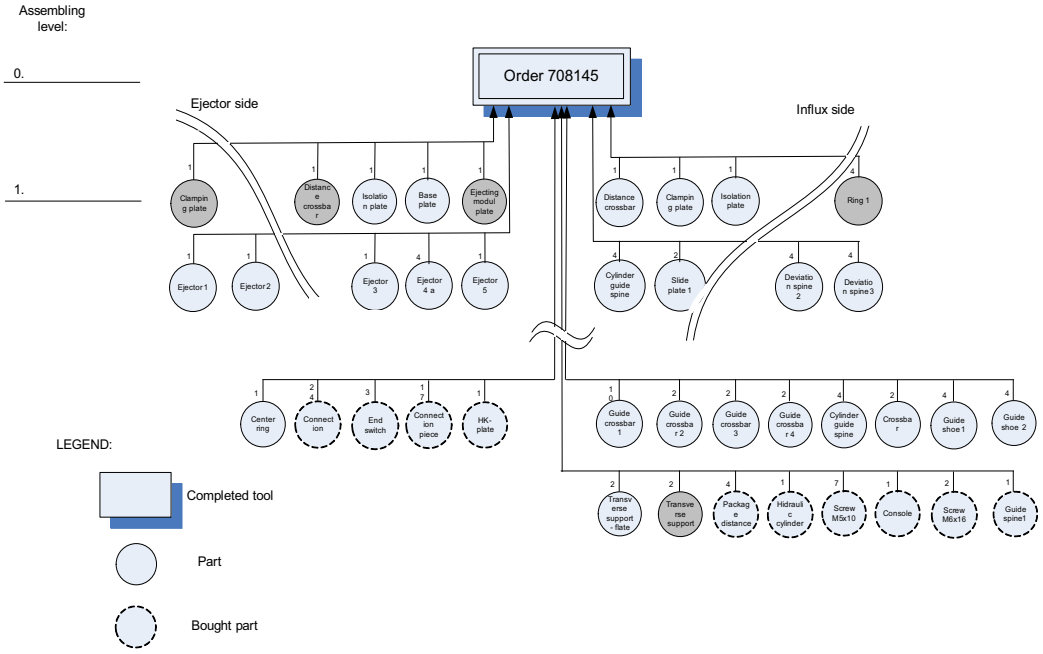


Fig. 15. Part of assembly structure of tool # 708145

As evident from Figure 15, the tool consists of two parts: ejecting and feeding parts. The tool consists of 122 parts (73 parts are manufactured in the tool shop and 49 parts are outsourced). There is just one assembly operation—the final assembly. After assembly, samples are manufactured and measurements are then taken.

Step 6. Establishing technology routings for manufacturing parts and assembly routings for assembling the components for the planned production order

The types and sequence of operations for tool # 708145 were used in this step (Figure 16).

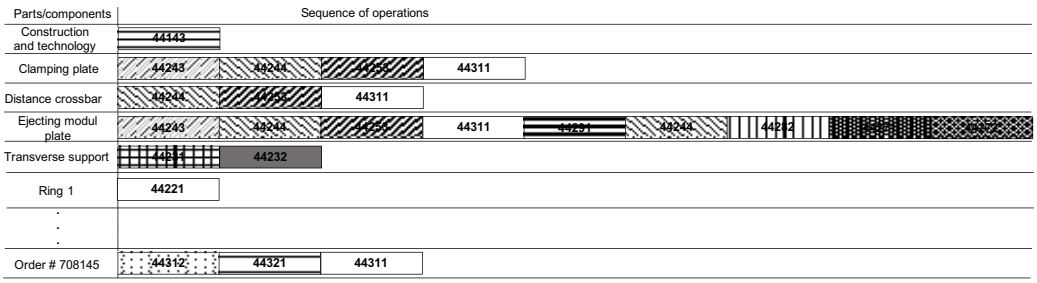


Fig. 16. Some types and sequence of operations required for parts and components of tool # 708145

For tool parts and components manufactured in the tool shop, it was necessary to obtain data on the type and sequence of operations, which ensure quality parts and components of

the tool. 231 operations were carried out on tool # 708145. Some types and the sequence of required operations on some parts and components of tool # 708145 are shown in Figure 16. This tool shop performs prior preparation for manufacturing; for this order it consists of: machine electronics (44142), design of tools (44143) and slitting (44211). The prior preparation itself could not be presented in the assembly structure in Figure 15; it does, however, increase the time for the order processing and must therefore be taken into account in operations.

Step 7. Random sampling of vector elements of forecast lead times of manufacturing and assembly orders for the planned production order – tool # 708145

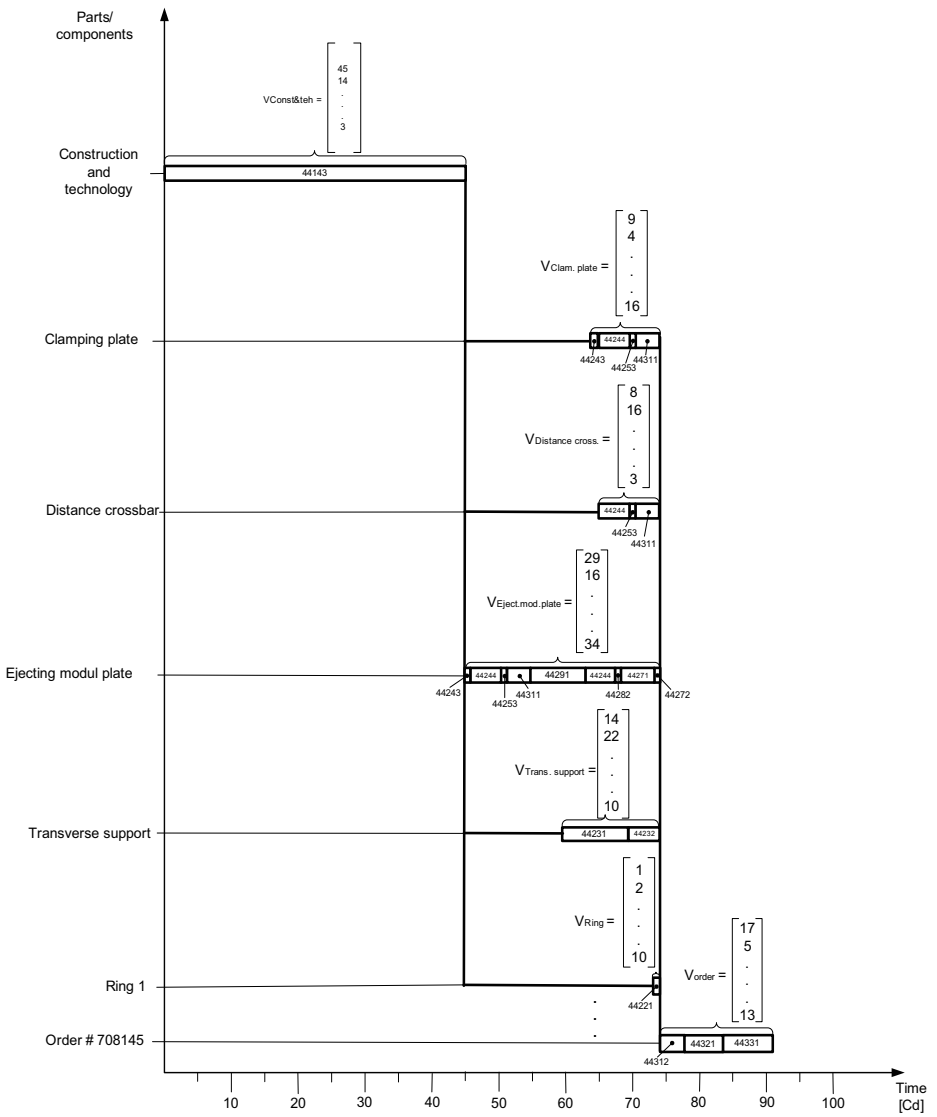
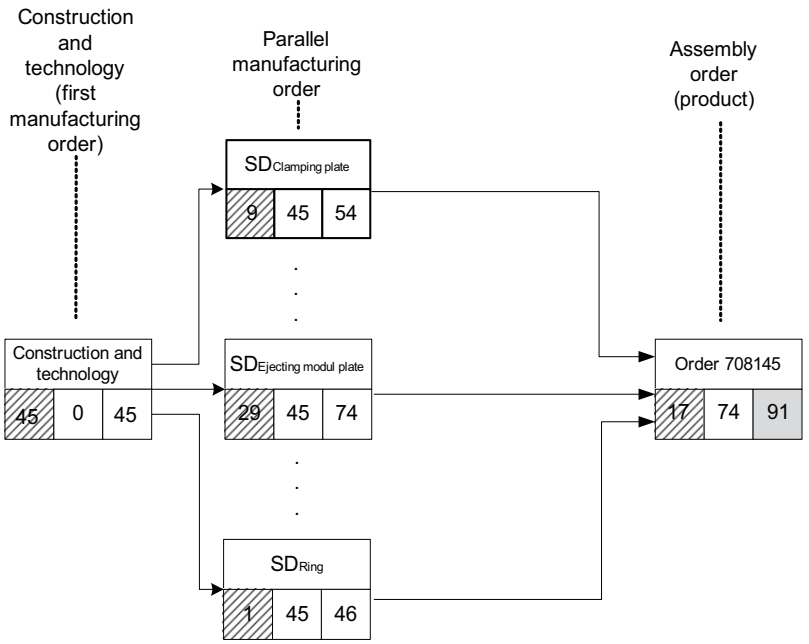


Fig. 17. Random sampling of vector elements of predicted lead times of manufacturing and assembly orders of the planned production order – tool # 708145

On the basis of the defined sequence of operations of parts and components of tool # 708145, Matlab software was used (The MathWorks, Inc., 2002) to form vectors of expected lead times of manufacturing and assembly orders, as described in the theoretical part of this paper.
On the basis of tests with 500, 1000, 5000, 10000, 20000 and 50000 samplings, it was concluded that 10000 samplings were enough for this case. If much fewer than 10000 samplings were used, forecasts between samplings differed considerably, because too few data were used. Using significantly more than 10000 samplings did not improve the result, it only increased the computing time.
On the basis of these 10000 samplings of lead times of operational orders, the vectors of expected lead times of manufacturing and assembly orders for tool # 708145 were defined (Figure 17).

Step 8. Forming a vector of forecast lead times of the planned production order – tool # 708145

In order to define the vector of expected lead times of the planned production order, 10000 samplings were made.
A sample calculation of lead time for the first sampling is shown in Figure 18. The expected lead time of the planned production order is the sum of the time of the first manufacturing order, maximum time of parallel manufacturing orders (parts) and time for the assembly order (final assembly of the tool).
After having completed 10000 samplings, a vector of expected lead times of the planned production order Vnar for tool # 708145 was obtained. Its lead-time-distribution function is shown in Figure 19.



$TO = TO_{Const\&teh} + \max(TO_{parallel\ manufacturing\ order}) + TO_{assembly\ order} = 45 + 29 + 17 = 91\ Cd$

Fig. 18. Activity network diagram for calculation of lead time for the first sampling of the production order # 708145

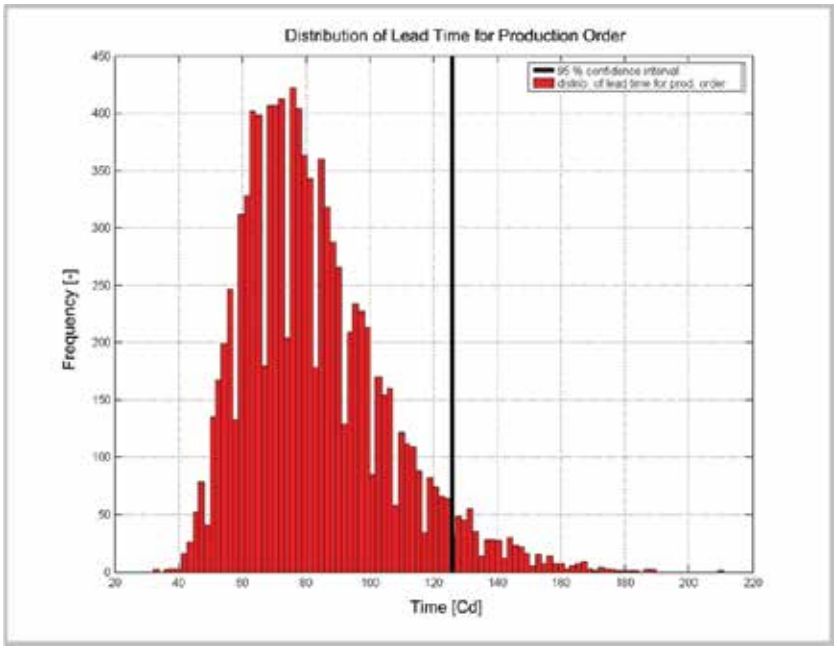


Fig. 19. Lead time distribution of the production order—tool # 708145

Step 9. Forecasting the delivery lead time of the planned production order – tool # 708145

No customer is interested in a lead-time vector of a production order (or its distribution function) as a delivery date, so the median TO_{med} of this vector is used as the first approximate value; for this order it is:

TO_{med} = 79 Cd.

The expected lead time for the production order is therefore equal to the 50th percentile of the V_{nar} vector of the expected lead times of the production order; so there is a 50% probability that the actual delivery time will be within the deadline, and 50% probability that it will not be within the deadline.

However, the median is not a sufficient estimate in engineering; instead, a 95% probability is required, which corresponds to the 95th percentile. For this production order the forecast lead time is: $TO_{95\%}$ = 126 Cd.

Ninety-five percent is a high enough forecasting probability, so it was suggested to the company that this value be used.

Each company must decide what risk level is acceptable for them when signing a contract with a customer.

As has already been mentioned, the company made bids in the past on the basis of experience and similar past projects. The proposed and signed deadline for tool # 708145 between the tool shop of Eti d.d. company and the customer was 73 Cd. The value of this project was 45,000 €. Penalties for delay were defined as 1% per week up to a maximum of 10% of the order value.

The order was delivered on 103 Cd (the delay of 30 Cd incurred a penalty of 1800 €). However, the penalty may not be the main problem—the main problem is that the tool shop in such a case loses its reputation as a good and reliable supplier—which is an invaluable asset.

Table 4 presents the deadline planned by experienced company employees compared with the actual deadline and forecast deadlines with various probabilities.

	Deadline [Cd]	Deviation from the actual delivery [Cd]	Deviation from the actual delivery [%]
Planned deadline	73	- 30	- 29,1
Actual deadline	103	0	0
Median	79	- 24	- 23,3
60% probability	85	- 18	- 17,5
80% probability	100	- 3	- 2,9
83% probability	103	0	0
90% probability	114	+ 11	+ 10,7
95% probability	126	+ 23	+ 22,3

Table 4. Comparison between the planned-, actual- and various predicted probabilities of order processing lead times

As evident from Table 4, the right forecast for this order was at a probability of 82%. At 95% probability (126 Cd) the forecast would exceed the actual date by 23 Cd.
According to the results obtained with many tests, it can be concluded that the procedure for forecasting lead times yields sufficiently accurate results.

6. Conclusion

Due to ever-fiercer market competition, and because of the transition from a seller's market to a buyer's market, companies must forecast lead times and delivery times with ever greater accuracy. If they give incorrect deadlines, they may not get a request from a particular company the next time, which can lead the company into crisis.
The article proposes a procedure for forecasting production-order lead times on the basis of actual lead times of past operational or assembly orders. Using the proposed procedure, the company can:
forecast the lead time required for delivery of any new order to any customer;
make variations of delivery lead time calculations on the basis of an acceptable risk level by selecting the confidence interval with respect to the size and complexity of an order, and taking into account the company's policy towards its customers. This means that the company can risk more to obtain an important order (narrower confidence interval). The company would thus have to prioritise this order during the manufacturing process, which may cause late delivery of other orders.
The procedure for forecasting lead times was tested several times and is presented in a case study of forecasting lead times for manufacturing a tool for a linking element of an oil vent in a tool shop in Slovenia. The case study was done using data gathered over the last three years in the Largo ERP system database.
Using this procedure, a sales department can make a well-defined bid for the customer in a short time. The sales person does not need many years of experience—(s)he only needs well-defined technology routings, while the company management provides a confidence interval. On the basis of these data, the delivery time for an order can be defined. The main

advantage of this procedure is therefore that companies will not depend so much on estimates made by experienced employees. They will use instead past data, stored in the ERP system or data manually recorded in the past.
On the basis of the tests, it was found that the procedure for forecasting lead times of production orders was well designed and provided very useful data for sales, as well as for production planning and control. Signing a supply contract on the basis of reliable statistical data is completely different from signing a contract on the basis of uncertain, experience-based guesswork.
It is planned that in the future the proposed procedure will be improved by taking into account the sequence of operations required to complete an order, the influence of the number of operations per order, and the influence of the processing time of operations.

7. Acknowledgements

We would like to thank to the tool shop company for giving us access to the data from their ERP system and for their technical aid. We would also like to thank to the Slovenian Ministry of Higher Education, Science and Technology for their financial aid during development of this method.

8. References

Begemann C., 2005, Terminorientierte Kapazitätssteuerung in der Fertigung, Dissertation Universität Hannover

Buzacott J.A., Mandelbaum M., 1985, Flexibility and Productivity in Manufacturing Systems, Proceedings of the IIE Fall Conference, Chicago, IL, pp. 404-413

Chen Y.J., Zhang M., Tseng M.M., 2005, An Integrated Process Planning and Production Scheduling Framework for Mass Customization, Proceedings of CIRP 3rd International Conference on Reconfigurable Manufacturing Systems

Denkena B., Lorenzen L.-E., Batino A. , 2006, Increased Production Flexibility and Efficiency through Integration of Process Planning and Production Control, Proceedings of the 39th CIRP International Seminar on Manufacturing Systems, pp.157-161

Ferligoj A., Osnove statistike na prosojnicah,Tiskarna Mišmaš, Ljubljana, 1995, pp. 26-28

Kingsman B.G., 2000, Modelling input-output workload control for dynamic capacity planning in production planning systems, International Journal of Production Economics 68, pp. 73-93

Kingsman B.G., Tatsiopoulos I.P., Hendry L.C., 1989, A structural methodology for managing manufacturing lead times in make-to-order companies, European journal of Operational Research 40, pp. 196-209

Krause F.L., Altmann C., 1991, Integration of CAPP and scheduling for FMS, Proceedings of IFIP—CAPE

Largo, 2007, http://www2.perftech.si/sxp/default.sxp?lang=1&wp=2&id=62&lid=2088

Lawrence S.R., 1994, Negotiating due dates between customers and producers, International Journal of Production Economics 37, pp. 127-138

Leem C.S., Suh J.W., 2005, Techniques in integrated development and implementation of enterprise information systems, Intelligent knowledge-based systems, Business and technology in the new millennium, Vol. 2, Information technology, Kluwer Academic Publishers, 3–26.

Lödding H., 2005, Verfahren der Fertigungssteuerung – Grundlagen, Beschreibung, Konfiguration, Springer-Verlag, Berlin Heidelberg New York

Nyhuis P., Vogel M., 2006, Adaptation of logistic operating curves to one-piece flow processes, International Journal of Productivity and Performance Management, Vol. 55, No.3/4, pp. 284-299

Nyhuis P., Wiendahl H.P., 1999, Logistische Kennlinien, Springer Verlag, Berlin, 81-94.

Rice J. A., 1995, Mathematical Statistics and Data Analysis, Second Edition, International Thomson Publishing, California

Scherer E., 2005, ERP – Projekte auf dem Prüfstand der Praxis, ERP Management, GITO mbH Verlag, 34-37

Starbek M., Grum J., 2000, Selection and implementation of a PPC system, Production planning & control, Vol. 11, pp.765-774

Tatsiopoulos I.P., Kingsman B.G., 1983, Lead time management, European journal of Operational Research 14, pp. 351-358

The MathWorks, Inc., 2002, Getting Started with MATLAB, ver. 6

Wang Z., Chen Y., Wang N., 2004, Research on Dynamic Process Planning System Considering Decision about Machines, Proceedings 5th World Congress on Intelligent Control and Automation, Hangzhou, China

Wiendahl H.P., 1995, Load-Oriented Manufacturing Control, Springer Verlag, Berlin, 37-199.

Wiendahl H.-P., Dammann M., 2006, Production Control of Products with Different Demand Volatility, Proceedings of the 39th CIRP International Seminar on Manufacturing Systems, pp.185-192

Ashjari B., Fatemi Ghom S. M. T., 2001, A heuristic method for resource constrained allocation in multi-project scheduling, Iranian Journal of Science and Technology Volume 26, pp.111 - 116

Begemann C., 2005, Terminorientierte Kapazitätssteuerung in der Fertigung, Dissertation Universität Hannover

Buzacott J.A., Mandelbaum M., 1985, Flexibility and Productivity in Manufacturing Systems, Proceedings of the IIE Fall Conference, Chicago, IL, pp. 404-413

Chen Y.J., Zhang M., Tseng M.M., 2005, An Integrated Process Planning and Production Scheduling Framework for Mass Customization, Proceedings of CIRP 3rd International Conference on Reconfigurable Manufacturing Systems

Denkena B., Lorenzen L.-E., Batino A. , 2006, Increased Production Flexibility and Efficiency through Integration of Process Planning and Production Control, Proceedings of the 39th CIRP International Seminar on Manufacturing Systems, pp.157-161

Enns S.T., 1994, Job shop lead time requirements under conditions of controlled delivery performance, European Journal of Operational Research 77, pp. 429-439

Fatemi Ghomi S. M. T., Torabi S. A., 2001, Extension on common cycle lot-size scheduling for multi-product, multi-stage arborescent flow-shop environment, Iranian Journal of Science and Technology Volume 26, pp. 55 - 68

Ferligoj A., Osnove statistike na prosojnicah,Tiskarna Mišmaš, Ljubljana, 1995, pp. 26-28

Kingsman B.G., 2000, Modelling input-output workload control for dynamic capacity planning in production planning systems, International Journal of Production Economics 68, pp. 73-93

Kingsman B.G., Tatsiopoulos I.P., Hendry L.C., 1989, A structural methodology for managing manufacturing lead times in make-to-order companies, European journal of Operational Research 40, pp. 196-209

Krause F.L., Altmann C., 1991, Integration of CAPP and scheduling for FMS, Proceedings of IFIP—CAPE

Kušar J., Brezovar A., Grum J., Starbek M., 2004, Realistic lead time scheduling of operations of orders. Int. j. mach. tools manuf.. [Print ed.], letn. 44, št. 10, str. 1037-1046.

Largo, 2007, http://www2.perftech.si/sxp/default.sxp?lang=1&wp=2&id=62&lid=2088

Lawrence S.R., 1994, Negotiating due dates between customers and producers, International Journal of Production Economics 37, pp. 127-138

Leem C.S., Suh J.W., 2005, Techniques in integrated development and implementation of enterprise information systems, Intelligent knowledge-based systems, Business and technology in the new millennium, Vol. 2, Information technology, Kluwer Academic Publishers, 3–26.

Lestan Z., Brezocnik M., Buchmeister B., Brezovnik S., Balic J., 2009, Solving the job-shop scheduling problem with a simple genetic algorithm, Int. j. of simulation modelling, Vol. 8, No. 4, pp. 197-205

Lödding H., 2005, Verfahren der Fertigungssteuerung – Grundlagen, Beschreibung, Konfiguration, Springer-Verlag, Berlin Heidelberg New York

Nyhuis P., Vogel M., 2006, Adaptation of logistic operating curves to one-piece flow processes, International Journal of Productivity and Performance Management, Vol. 55, No.3/4, pp. 284-299

Nyhuis P., Wiendahl H.P., 1999, Logistische Kennlinien, Springer Verlag, Berlin, 81-94.

Öztürk A., Kayaligil S., Özdemirel N.E., 2006, European Journal of Operational Research 173, pp. 683-700

Rice J. A., 1995, Mathematical Statistics and Data Analysis, Second Edition, International Thomson Publishing, California

Scherer E., 2005, ERP – Projekte auf dem Prüfstand der Praxis, ERP Management, GITO mbH Verlag, 34-37

Starbek M., Grum J., 2000, Selection and implementation of a PPC system, Production planning & control, Vol. 11, pp.765-774

Tasic T., Buchmeister B., Acko B., 2007, The development of advanced methods for scheduling production processes, SV – Journal of Mechanical Engineering, Vol. 53, No. 12, pp. 844-857

Tatsiopoulos I.P., Kingsman B.G., 1983, Lead time management, European journal of Operational Research 14, pp. 351-358

The MathWorks, Inc., 2002, Getting Started with MATLAB, ver. 6

van Ooijen H.P.G., Bertrand J.W.M., 2001, Economic due-date setting in job-shops based on routing and workload dependent flow time distribution functions, International Journal of Production Economics 74, pp. 261-268

Vig M.M., Dooley K.J., 1991, Dynamic rules for due-date assignment, International Journal of Production Research, Vol. 29, No.7, pp. 1361-1377

Wang Z., Chen Y., Wang N., 2004, Research on Dynamic Process Planning System Considering Decision about Machines, Proceedings 5th World Congress on Intelligent Control and Automation, Hangzhou, China

Weeks J.K., 1979, A simulation study of forecastable due-dates, Management science, Vol. 25, No.4, pp. 363-373

Wiendahl H.P., 1995, Load-Oriented Manufacturing Control, Springer Verlag, Berlin, 37-199.

Wiendahl H.-P., Dammann M., 2006, Production Control of Products with Different Demand Volatility, Proceedings of the 39th CIRP International Seminar on Manufacturing Systems, pp.185-192

8

The Market for NPD Services: the Emerging Business Models in Italy

Valentina Lazzarotti and Emanuele Pizzurno
Università Carlo Cattaneo – LIUC
Italy

1. Introduction

Over the last few years, the increasing market turbulence and competition, the time reduction of products life cycle and the increased variety of products and services have required companies to be more flexible in order to respond effectively to market requirements. In today global competitive environment, one of most important decision that managers deal with is the innovation and developing of the technology: inside o trough external organisations? In this perspective, nowadays companies aim at developing only their own core competences and to outsource no-core activities, even commonly considered strategic, like R&D (Quinn, 1999). According to this, a "market for technology" is growing (Arora et al., 2001), in which companies offer (a set of) services supporting the R&D process of other companies. Within this framework of studies an increasing trend in outsourcing activities linked to the new product development have been explained (Zhao & Calantone, 2003). The growing relevance of the firms offering NPD services is confirmed within the emerging literature on the R&D management. This importance arises from the widely established trend in outsourcing different (part of) activities concerning the innovation process and of the R&D function, even if considered, up to today, as strategic (Emes et al., 2005). Past studies have extensively pointed out that through this externalization, several companies gain different competitive advantages (higher quality, lower costs, faster-to-market products, etc.).
However, most of the available contributions are more focalised on factors which can explain the reasons and benefits to access to external sources for innovation and less on the suppliers of these services. In fact, previous works have shown as the effectiveness of new product development requires the support of external. Thus, referring to this well developed context, a market of technology has increased (Arora et al., 2001). A large number of firms offer a wide set of "knowledge intense (business) services in which, the NPD services are playing a relevant role. Many researchers have examined and confirmed that this market is really growing in terms of number and size of enterprises (Chiesa et al., 2004). Recent studies have showed that the light on these firms has been shed, focalising on the industrialised countries (Chiesa et al., 2004) and referring to specific country peculiarities (Veugelers and Cassiman, 1999) or focusing on a specific firm – especially on the American IDEO – (Hargadon and Sutton, 1997; Kelley, 2001; Chiesa et al., 2004). What it is still not studied, in depth and in a wider sample, it is the way these firms are managed (in terms of structure, organisation, services provided, and network of collaborations, etc.).

More precisely, seem to lack suggestions on how the elements of the business model should be composed in a coherent set of relationships to achieve effective and efficient NPD service. Thus it is this stream of works we attempt to help with this paper by deepening the study of the organizational and managerial features for a sample of service Italian companies. In particular, what we thought it was interesting to delve mainly into the case of companies that declare to support the entire process of new product development. If compared to companies offering single stage services, such companies face in fact a high complexity, given that they have to turn an abstract idea into something that is "real", "concrete", and ready to be sold. This means that they have to solve not only the problems of each phase of the NPD process, but also those relating to (i) relationships among the various phases of development, (ii) many competencies and resources needed, (iii) interaction with the client during the whole process (Emes et al., 2005; Jurgens, 2000; Vegulers & Cassiman, 2003). As a consequence, the picture that can be drawn analysing such companies provide interesting suggestions because it allows to appreciate the relationships among the several business model elements (and how they should be composed in order to provide a certain set of NPD services, thus giving concrete managerial suggestions in the sense clarified above). Moreover, a similar result can also benefit companies that offer just one or few stages of the process: they should obtain suggestions in order to decline in a different way their specific business model.
Coherently with the paper goal, the empirical study presented here is a multiple case study, aimed at describing how companies offering services for the NPD process organise and manage their business. Twenty-five companies have been studied through interviews conducted with all the top – or project – managers, trough a semi-structured questionnaire.
The empirical study has been conducted by a twofold aim:

- to search for TSS companies that support the entire process of new product development and/or, conversely, that provide services to a smaller part of the innovation process or only on few stages (the other extreme, i.e. only one phase) in order to identify clusters of companies. In other words, we try to identify clusters by using the element in the companies' business model representing the "completeness of the service" offered (in a continuum that goes from one stage to the whole process);
- once identified clusters, to study if they show differences in terms of the other investigated organisational and managerial features (all the other company's business model elements). In other words, we consider these other variables as illustrative ones in order to describe each emerged cluster.

The analysis of the gathered data has actually allowed to identify some typical profiles of TSS companies in which the elements of the business model take a particular and coherent combination. These clusters are described in detail in the paper, which is organised into four different sections:

- description of the conceptual context of this study, giving the basic theoretical background, concepts and definitions;
- research methodology;
- empirical study: description of the case studies and analysis of data gathered;
- conclusions and future research.

2. The conceptual context of the study: the KIBS and the TSS within the new product development process

The purpose of this section is to briefly introduce some concepts and definitions useful to understand what our field of empirical investigation is.

Several authors have studied and demonstrated that, in the innovation development field also, interaction with external entities is growing (Tidd, 1995; Quinn, 1999, 2000; Chiesa et. al., 2004; Lazzarotti & Manzini, 2009). This tendency towards outsourcing innovation has created a new category of services called KIS – knowledge intensive services (Windrum & Tomlison, 1999) – or KIBS – knowledge intensive business services (Miles, 2000; Muller & Zenker, 2001; Knoben & Oerlemans, 2006; Strambach, 2008; Horgos & Koch, 2008; Zenker & Doloreux, 2008) – characterised by a high innovative level and scientific intensity of the outputs. According to Windrum & Tomlison, 1999, 'private sector organisations that rely on professional knowledge or expertise relating to a specific technical or function domain. KIS firms may be primary sources of information and knowledge or else their services form key intermediate inputs in the products or production process of other businesses.' This kind of service can be applied to several sectors: from banking to real estate, from market research to insurance services. Among KIBS, a more specific subset can be identified, called TSS – Technical and Scientific Services (Abetti, 1989; MacPherson, 1997a; MacPherson, 1997b; Howells, 1999; Larsen, 2000; Chiesa & Manzini, 2001; Arora et al., 2001; Chiesa et al., 2008; Chiaroni et al., 2008). According to these authors, TSS are "services which rely upon technical and scientific knowledge and give an output that is, again, technical and scientific knowledge." In other words, they are service companies that sell technology and scientific knowledge. What joins these companies and differentiates them from others that fall within the definition of KIBS is therefore the nature of knowledge on which they are based and they incorporate into their services, that is technological.

Literature (Windrum & Tomlinson, 1999; Debackere 1999; Chiesa & Manzini, 2001, Chiesa et al., 2008) has also suggested several taxonomies in order to identify existing service firms: for instance, companies are grouped by type of the output provided (e.g. work-in-progress innovation that is an intermediate finding that needs to be further developed to be commercialized as an innovation; single process activity, that means TSS firms carry out, for the client company, a stage of its innovation process; whole new process development process, that means service companies start from an idea and provide their client with a new product ready to be put into production and then commercialized; technologies to develop technologies, in the case that TSS provide technologies that can be used in order to improve the efficiency and the effectiveness of the client's company's innovation process); technical and scientific competences, the TSS firm is based on and incorporates in its services, i.e. the technical or scientific domains in which an excellent knowledge level has been reached (e.g. mechanical engineering, genomic, microchip design); the client firm's sector of activity (e.g. mechanic, electronic, chemical, pharmaceutical); the generality of the output provided by the TSS company, a general service, i.e. aimed at supporting the innovation process of firms from different sectors or a specific service, i.e. addressed to innovative firms of a specific sector. From all these definitions and taxonomies emerge nonetheless clear that TSS services are important supports to the companies' technological innovation process.

For the purposes of this study, however, we take a narrower perspective and focus particularly on new product development services within the broader process of technological innovation. To this end a recent classification (Chiesa et al., 2008) can help us. Firstly, it considers as relevant dimension the generality of the output provided by the TSS company, as just defined above. Regarding this dimension, we consider here the generic services without choosing any sector specialization. This choice is aimed to avoid results that are significantly biased by industry-specific factors.

Secondly, supported from the wide literature on the subject, it allows us to define the new product development services within the innovation process: four categories of product development activities, highlighted in bold, where companies, which we will study in this paper, are involved.
More in detail, among the many existing contributions that provide their own versions of the NPD process and related activities (here we just mention the most important, such as Urban & Hauser, 1993; Cooper, 1994; Kotler, 1997; Jones & Stevens, 1999; Haden et al., 2004; Rundquist & Chibba, 2004; Varela & Benito, 2005; Cooper, 2008; for a review see Trott, 2008), we follow here basically Verganti's definition (1997) that identifies the detailed tasks composing the first three groups of relevant activities, to which we add the phase of launch and commercialization (this one, according to Kotler, 1997). In this regard, see table 1 where the detailed list of activities is shown.

PHASE	ACTIVITIES
Concept generation	Definition of briefs Analysis of customer needs Competition analysis Definition of the generic product Generation, testing and selection of concept Assessment of the investment in new product Formulation of project plan
Product design	Design - architectural adaptation Choice of technologies and components Choices to make or buy design Definition of the detailed specifications of those components Design modules-components Prototyping and testing of component quality Integration of modules-components Test Product Quality Optimization
Engineering (process design)	Configuration of the production process Design of machines and tools (dies, tools etc.) Development of part programs to control production machines Definition of schedules and work instructions Definition timing and methods and workforce training Pre-implementation Start production
Launch and commercialization	Launch the product Produce and place advertisements and other promotions Fill the distribution pipeline with the product Pricing

Table 1. List of relevant NPD activities

In summary, in this work we focus on services TSS intended as new product development services that support one or more stages of development of an innovative product without having a specialization addressed to a specific industry. At this point we have all the

elements to study the selected companies that provide these services, after a brief clarification about the adopted methodology.

3. Research methodology

The research method adopted in this work is based on a multiple case study. Despite the widely acknowledge limitations of this approach, especially in terms of reliability and validity (Ginsberg & Abrahamson, 1991; Yin, 2003), the case study method has the ability to capture the full complexity of the studied phenomenon, including its 'softer' aspects. Given that the aim of our empirical study was to investigate TSS practices in-depth, the aforementioned advantage of the case study method was a critical factor in selecting the research approach.

Information was collected through direct interviews with companies' management and internal documents were also consulted. The applied research methodology had two main limitations. First, because the empirical base was mainly built up from personal direct interviews with the company's top manager, the results are susceptible to bias arising from distorted and subjective interpretations and rationalizations. Second, as in most case studies, the empirical research does not permit any systematic generalisation. That said, the aim of this empirical investigation was not to generalise, but rather to offer a detailed description of the phenomenon and to offer some new insights for future investigations, aimed at generalising results (Eisenhardt, 1989).

4. The empirical study

4.1 The sample

The empirical study presented here is a multiple case study, aimed at describing how companies offering services for the whole NPD process organise and manage their business. Limiting the search to a first step, we decide to start by focusing on a population – and consequently selecting a sample – consistent with these criteria:

- firms that have declared able to support the entire product development process;
- firms that are not specialised in supporting NPD within a specific industry;
- private-owned firms that have NPD services as core business;
- firms able to develop a physical finished good and, among the sectors, excluding those that show specific scientific peculiarities in NPD processes (Trott, 2008) i.e. software, pharmaceutical, chemical or biotechnological companies.

Twenty-five companies with these features have been studied, as reported in table 2 through:

- interviews: more than 3 telephone and in-person interviews were conducted with all the top – or project – managers, trough a semi-structured questionnaire. The questionnaire is too wide to be included in this paper; however, respondents were asked questions related to:
  - The NPD company organization as: (i) firm organisation and services offered, (ii) HR management, (iii) knowledge management, (iv) firm structure, (v) competencies and collaborations;
  - The commercialisation and internationalisation strategy intended as: (i) sale of NPD services and external communication, (ii) CRM (iii) clients' business sector of activity, (iv) client searching, (v) competitors, (vi) location of clients and (vii) pricing;

  - The project organisation and management, in terms of (i) interaction with the client during the NPD process, (ii) commercialisation phase, (iii) intellectual property management, (iv) performance measurement system, (v) projects average duration, (vi) project management and (vii) inter-functional teams.
- documents, both internal (provided by interviewed people, such as internal project reports and prototypes) and public (available, for example, on the web-sites of companies, such as presentations and promotions).

Then, a structured cross – case analysis was carried out, through which data and information collected have been elaborated, categorised and compared in order to point out analogies and differences, so as to draw a reliable and synthetic picture of the sample analysed.

	Company	**Revenue**	**Employees**
1	Appliances Engineering	1 – 5 mln €	11 – 20
2	Attivo Creative Resource	200.000 – 1 mln €	1 – 10
3	Creanova	200.000 – 1 mln €	1 – 10
4	Design Continuum	confidential	11 – 20
5	Design Group Italia	1 – 5 mln €	11 – 20
6	Disegno Bello	200.000 – 1 mln €	1 – 10
7	DNA	confidential	1 – 10
8	Esseti	confidential	1 – 10
9	Far Design	< 200.000 €	1 – 10
10	Fox Bit	1 – 5 mln €	51 – 100
11	James Irvine	200.000 – 1 mln €	1 – 10
12	MR&D Institute	> 5 mln €	51 – 100
13	Partec	< 200.000 €	1 – 10
14	Pininfarina Extra	> 5 mln €	21 – 50
15	Pro Design Italia	confidential	1 – 10
16	Promau	1 – 5 mln €	51 – 100
17	SB3	1 – 5 mln €	1 – 10
18	Sintesi AB	200.000 – 1 mln €	1 – 10
19	Sowden Design	200.000 – 1 mln €	1 – 10
20	Spring Design	confidential	1 – 10
21	Studio Bonfanti	200.000 – 1mln €	1 – 10
22	Studio Primalinea	200.000 – 1mln €	1 – 10
23	SZ Design (Zagato)	> 5 mln €	51 – 100
24	Vegni Design	confidential	1 – 10
25	VIP Techologies	< 200.000 €	1 – 10

Table 2. The companies studied

Finally, the main evidence and findings emerging have been discussed with some of the people interviewed, in order to verify their validity. Only one firm has been excluded by the analysis because clearly emerged - during the empirical research - as unable to offer a various set of NPD services. The main conclusions of the study are presented in the following section.

4.2 Research finding: the Italian NPD firms

In this paragraph, the organizational and managerial features of the studied Italian NPD firms are described. First of all, the main analogies are analysed, i.e. those elements that characterise in a very similar way all the companies studied; then, the significant differences are pointed into evidence, i.e. those elements in the business model that significantly diverge among companies, with the aim to verify whether some clusters of companies can be identified.

4.2.1 Analogies among the Italian NPD firms

Firstly, analogies are described in terms of: company's organisation, commercialisation and internationalisation strategy, project organisation and management.

Company organisation

Firm structure: while the firms are showing a different organisation in relation to their dimension (in term of employees), all of them – even the smallest – follow a common approach: the matrix management and, more often, a strong (project) matrix. A project manager - who is primarily responsible for the project - is always identified. Functional managers provide technical expertise and assign resources on an as-needed basis.

Human resources management: this can considered a key point for the success of the NPD firms, which base largely their activities on the competencies and experiences of their employees. In particular, the critical factors in HRM for NPD companies are: (i) the recruitment of talents or well-trained personnel; (ii) the continuous improvement of the capabilities; (iii) job environment and team working. Significant time and resources are dedicated to improving performance in these three topics. In terms of profile, usually graduates in Economics, Engineering (mechanical, electronic, management…) and Industrial Design can be found, but also Design licentiates. A significant percentage of employees are very young.

Knowledge management: even if the formal storage of past projects results and solutions is recognised to be effective and efficient, the Italian NPD firms usually do not use sophisticated tools to this aim. Even when designed and realized, these archives are not easily accessible and commonly used. Informal and personal relationships and networks generally represent the most important KM system, together with the storage of prototypes and pictures of past developed products.

Commercialisation and internationalisation strategy

Sale of NPD services and external communication: the market of services for NPD is still unknown, even if the externalisation of NPD seems to be increasingly relevant. Main consequence of this lack of knowledge, shared language and sufficiently precise and widespread classifications is that the marketing of the services offered is a very difficult task for NPD service companies, which have to rely on their own capability of self-introduction to the market. More precisely, it has been recognised that it is very difficult for NPD

companies to clearly communicate what they are really able to do. Web-sites and visiting professional fairs represent the main channels for the external communication. As a consequence, information given on the web sites, if used, tries to be rich, detailed and well structured. A section, called "credentials" is frequently used to better clarify the company's activity and qualification, in which previous projects of new products are described. This difficulty in communication is also recognised as the main barrier in the acquisition of new clients, together with the problem of evaluating the "value" of the services offered.
Customer relationship management: the Italian NPD companies show, on average, 80% of continuative relations against a 20% of single projects committed. All main CRM techniques are well known and widely used.
Prices and margins: the complexity of the project determines the price (from few thousands Euros to hundred thousands Euro); the average is around 80.000 € - 100.000 €. In price definition the following parameters are taken in consideration:

- time (in term of hours – or days/weeks – required to perform the project);
- kind of activities performed. In fact not all activities have same evaluation (mechanical design is considered less than electronic design, for instance);
- kind of resources employed (human, technical, etc.);
- estimated value of the product on the market .

The average expected margin on the projects is around 20%.

Project organisation and management

Interaction with the client: the level of computer-aided support reached in the last years, above all concerning innovative solutions for remote collaboration, is extremely advanced. Even if software and ICT tools for remote collaboration are well known by companies, they are far from being diffused and widespread adopted. All NPD firms agree that the main cause reside in the client culture, which suggests adopting a more traditional approach. Companies still prefer personal relationships in NPD. As a matter of fact, even companies with a size and a level of project complexity totally adequate to the use of remote ICT tools, base their co-operation with clients upon:

- periodical meetings/contacts with the customer in order to define the concept (through the due brainstorming), to consult together with the technicians and to present the work in progress and the finished project;
- traditional communication tools, such as fax and e-mail.

Also within the NPD firm itself, such radical changes are not undertaken, and traditional updates and periodical meetings among members of the inter-functional team take place.
Commercialisation phase: this phase of the NDP process is strictly controlled by the customers and it is never outsourced to NPD firms, even when such firms offer adequate competences and skills. It is quite evident that this is probably the most critical phase for clients to ensure the appropriability of their innovation.
Intellectual property management: in all the observed companies, by contract, the intellectual property of the new product is owned by the clients. The patenting process is usually outsourced to specialists and considered as a standard service for the clients.
Performance measurement system: typically, the performance measurement system (PMS) in NPD projects takes into consideration conventional economic performance indicators (costs, timing, quality, resources, clients satisfaction, level of sales of the new product compared to forecasting, etc.). In several NPD firms, it has been observed an increasing diffusion of approaches to the measurement of innovative performances (i.e. commitment

and creativity of employees, new clients and other qualitative indicators as company reputation) aimed at monitoring the company's innovative capability, processes and results (Chiesa et al., 2006).
Project management: a project manager – who is primarily responsible for the project – is always identified. Functional managers provide technical expertise and assign resources on an as-needed basis. Basic techniques of project management are well known and widely used. Project teams are created on the basis of adequate skills and knowledge of the specific customer. The project manager and the team use the most diffused tools, software and methodologies in the technical-design area (concurrent engineering, Design for X, etc) as well as in the managerial one (forward engineering, WBS, milestones, Gantt diagrams, etc). The project teams usually interact with the project's client in correspondence of the defined milestones, when the state of the art is verified, the possible delays are defined and the needed corrective actions are identified, and all information about the project is shared. These meetings are also critical for evaluating qualitative "soft" factors, such as the development of experiences, in a business collaborative atmosphere.

4.2.2 Differences among the Italian NPD firms

Secondly, differences are here analysed in terms of company's organisation, internationalisation and commercialisation strategy.

Company organisation

Firm organisation, services offered and related competences: size is significantly different among the studied firms. Companies can be constituted by a limited staff – from 1 to 10 employees - or be structured with highly remarkable resources (from 51 to 100 employees). In the first case the firm structure results extremely flat, with owners and employees to fulfil basically the same tasks. In small companies, few employees manage the entire project, occasionally creating inter-functional teams with client's people. Big companies are more structured and present a formalised structure, with the following functions (that correspond to services offered to the customer and related competences, along the NPD process):

- Marketing: this function is dedicated to strategic marketing and it is able to carry out any qualitative and quantitative analysis (trend definition, product placement, market share calculation, etc.);
- Industrial Design (ID): this function generally works in coordination with the Mechanical and the Electronics Design departments in order to merge and coherently integrate the various ID objectives (such as strategic design, product and graphic design and brand development, ergonomics, functionality of shapes, materials, …). This function/unit usually involves very specialised personnel, it uses state of the art knowledge and technology and, hence, it provide a high quality outcome, integrated with the technical solutions adopted. It should be noted that some firms tend to specialize themselves on the more technical aspects of the industrial design. In these "technical oriented" firms, ID service is thus only partially offered, since aspects as brand development or aesthetic design are neglected.
- Mechanical Design: Mechanical Design is usually performed at excellent levels. New technologies are constantly adopted, with the aim to deliver new solutions that better deal with human factors (i.e. specific needs and constrains that derive from the direct interaction among human beings, products and technology) and with the growing sensitiveness to aesthetics.

- Electronics and Software Development: very few companies are actually able to offer this part of the NPD services according to the state of the art knowledge and technology. Indeed, some firms completely lack such competencies. As a consequence, when a new product requires electronic or SW subparts, these firms acquire outside a "shelf solution" or, in some case, collaborate with external sources (in some case with the clients themselves). Also "technical oriented" firms (which focus mostly on the technical aspects of the NPD, as defined above) has proved very partial experiences in this field.
- Engineering: the supply of these services implies for significant investments. Due to the high rate by which the necessary assets become obsolete and the high investments related, only the most developed firms can afford pre-production and rapid-prototyping machinery. Moreover, this set of skills usually includes laboratory test and analysis, production of simple prototypes and supplier selection, competencies normally diffused in all the studied companies.

Competencies and collaborations: the mix of internal/external competencies is extremely various. In big firms the access to external sources of knowledge and technology is very low (just rapid prototyping or software development); these companies relies only on internal workforce. If the case, the external partnerships are arranged with a twofold aim. Sometimes, the external collaborations are stable and they cover the range of services offered by the NPD firm that, in this way, can enlarge and enrich the specialised internal competencies. In other cases – especially where a wide range of services is performed by a considerable internal staff – the occasional external collaborations are useful to support NPD company in periods of intense work.

These collaborators are often professionals or small firms specialised in one phase of NPD process. Thus, if NPD firms can have a maximum of three external established collaborations, other NPD firms can have tens of external partnerships. Rarely, the collaboration involves universities. Usually, partners are involved with different types of collaborations (and consequently with different levels of coordination and integration), for instance: (i) as members of the team developing the new product, or (ii) as a simple suppliers.

Commercialisation and internationalisation strategy

Clients' business sectors: all companies support NPD process in several different industries and it has not been noted an association between specific products and specific group of NPD firms (the client's business sector crosses among firms' groups). Automotive, telecommunication and electronics are the most diffused (more than 70% of NPD firms have clients belonging to these industries).

Location of the client: NPD firms are positioned close to the national economic neuralgic centres. For these companies, the location in an economically important area is important also to be close to potential customers. Furthermore, it allows a NPD company to be better and faster informed about their working field: conferences, seminaries and other similar activities are held usually in such economic centres. Some NPD companies serve clients in Europe and worldwide, above all the biggest ones. Anyway, interviews proved that clients' location rarely causes managing problems for Italian NPD firms, due to the availability of electronic tools. However, the physical distance can cause problems: the nature of the activities (i.e. creative ones) leads to the need of a direct contact with its own customers.

The internationalisation strategy: the internationalisation strategy of companies seems to be a critical point, since globalisation allows to significantly widening the potential market. Having a geographically wide market immediately points out the "distance" problem. Together with the "physical" problem described above, also cultural, legislative and linguistic differences are factors that tend to undertake growing importance, and above all when the concern different continents. Furthermore, an international market seems to be accessible only by big service firms. As a matter of fact, large dimensions seem to attract better skills, which, in turn, influence the diffusion of the firm's work. Generally, it has been observed that companies with similar skills but different dimensions (small vs. big) cover different sizes of markets, where the main distinction is between intra (i.e. small companies as Fardesign, Bonfanti) and inter-continental ones (big companies such as Design Group Italia, Attivo Creative Resource).

When the intercontinental market is becoming consistent, local needs start to emerge and thus companies have to be open new offices abroad (as in the case of company "MR&D Institute"). Physical proximity, obtained through new seats, will answer the aforementioned needs of filling cultural, linguistic and legislative gaps. De facto, new seats are never as big and structured as the headquarters and do not have all the skills of the latter either. They focus mainly on marketing, market analysis and design activities - that is on those factors that are more distance-sensitive - while the more technical phases of each project will be forwarded to the head office (Design Group is an example in this sense).

From this point of view, another factor seems to affect the internationalisation strategy: the firms' main focus of activity, i.e. the phase, within the NPD process, to which the company dedicates its main resources and/or those in which it is considered as excellent:

- An orientation to the "soft" design aspects (i.e. style, elegance, interface innovation) of NPD seems to experience a wider geographical feedback than technical ones (for example, this is the case of Pininfarina Extra and Zagato, that have an image and a brand appreciated worldwide).
- In contrast, an orientation to technical contents tends to become anonymous, as often lacking of a proper image. Without the necessity of looking for a certain line or style, the attractiveness for customers to face more expensive and difficult collaborations with distant partners seems to disappear (as declared by Pro Design and VIP Technologies). This, of course, limits the possibility for NPD companies of offering their services to geographically far companies. Moreover, the very technical part of the project is often developed in collaboration with the client itself, due to the fact that technical know-how of its products is hardly surpassed by that of the NPD firm. Thus, normally, customers are looking for a geographically closer partner rather than one which is further away.

In conclusion, more design-oriented firms will thus tend to have a more geographically heterogeneous market, while Technical oriented ones will basically work with national or even regional clients.

Clients searching: Shortly, the main potential channels for customer retrieval are: brand importance (of the NPD firm), personal acquaintances, word of mouth, marketing/trade dedicated employee, Web site and presence at professional fairs. Theoretically, each kind of NPD firm could define its specific mix of the aforesaid channels, thus distinguishing from the others. Practically, the importance of the personal acquaintances and word of mouth channels is a common point between all kinds of NPD firms, although Web mode seems more diffused for companies providing the entire set of NPD services.

Competitors: firms able to offer a very complete set of NPD services face a strength international competition (IDEO, Well Design, Cambridge Consultants, are the most famous competitors). On the other hand, firms which are specialised in few phases compete mainly on a local basis, with very small firms and professionals.

Project organisation and management

Project average duration: in general what influences the duration of a project are the technical (mechanical and electrical …) components and phases more than design ones, while the overall length is related to the complexity of the project and, mostly, to the number of phases covered. Normally, firms that cover the entire NPD process have an average one-year duration, period that is reduced to 2 months in firms covering only few phases.

Inter-functional team: the inter-functional teams within an NPD firm are a direct consequence of its structure and, hence, of its dimension; it is verified by the information gathered, how:

- smaller companies (maximum 4-5 employees) carry forward projects involving a single internal person - or a couple at most - and organising at times inter-functional teams together with part of its client's staff (Fardesign, Bonfanti, Gloss Design). Also firms constituted by a range of 6-10 people, work in a similar way, having 2-people-team supervised usually by one of the company managers or seniors (Attivo Creative Resource, Studio Prima Linea). This is a consequence of small firm dimensions and of homogeneous internal skills. In fact, at these levels, the company structure is very flat and similar skills are present, which obstacles the creation of structured teams. Project complexity is usually coherent with these firms' capabilities, and excessively long and complicated developments are carried forward together with the customer's personnel.
- Differently, it is with a more consistent, structured and heterogeneous staff that specific-competence-mixed internal workgroups start to be observed. Anyway, joint-work together with the client does not disappear in these cases (MR&D Institute, Design Continuum).

4.3 The emerging business models of NPD companies

All of the interviewed companies have described themselves more or less explicitly, through words and images, as capable of performing any necessary activity to create a functioning product out of an idea. In order to completely develop a new product an NPD firm should at least be skilled in the market analysis, industrial design, mechanical, electric and electronic design and software development areas. But only few companies "realized" the described profile, while the rest cover just a part of the complete development process. The most widespread missing skills of such companies are the electric/electronic/software development ones. This means they will be able to conceive and design a certain variety of products, but they are not able to complete (nor autonomously nor through their nets of collaborators) the development of an object that would require electronic or software applications. Lacking of one or more skills to be considered a complete NPD firm, we suggest naming this set of companies "Integrated Industrial" according to their own high-level industrial design competences.

The results of the empirical analysis showed that the considered market is rather fragmented in terms of service offered. However, some clusters can be identified. They can be represented as in Figure 1, considering at first the relationship between the completeness

of the service provided (i.e. all the phases of NPD are supported) and the level of one the managerial variable studied above (i.e. the level of internal competences). They are named as follows:

- Complete new product development firms (identified in the Figure 1 with the letter "C"): companies able to plan and develop a complex product, providing a high level of novelty and supporting the client in all the phases of the NPD process and to offer strategic consulting as well;
- Integrated industrial design firms (ID+) type I: companies which possess competencies in all NPD fields, except for software and electronic or electro-technical functions. In some cases all these competencies are internal, in other cases they rely on an effective external network of partners;
- Integrated industrial design firms (ID+) type II: however they present themselves as able to offer services as in the previous group, these small companies tend to be specialised in one (or more) phase of the NPD because they rely only on internal competencies; in consequence services offered is decreasing whether internal resources decrease.

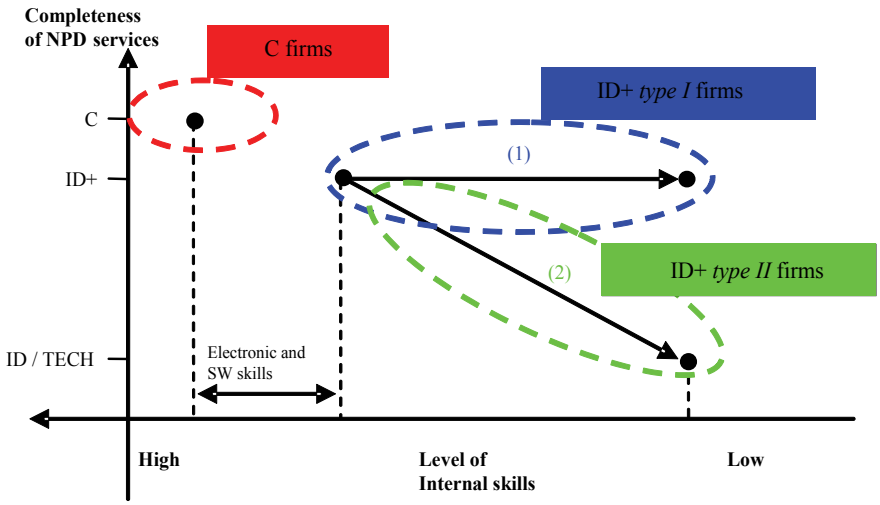


Fig. 1. NPD emerging clusters and business models

Thus, we found that the main difference in terms of competences is concerning the electronic and software skills. The figure can also be interpreted in a dynamical sense and it may represent a sort of path that companies can follow to increase the completeness of the service. In fact it highlights how companies with more expertise are able to offer a more complete service (cluster C); thanks to the use of external expertise, along with those inside, companies are able to maintain the completeness of the service (type I); instead, with the decreasing of internal expertise, if companies do not resort to external competences, completeness of service decreases significantly (type II). In the studied sample all companies started their history as "type II". Some of them, after a number of years, having observed growing requests by clients in different areas of knowledge, began to fill missing competences through the network of collaborations, transforming them into ID+ type I. If these requests started to be established these companies internalized all collaborators,

adding electronic and software know-how, transforming into C NPD firms. Secondly, Table 3 summarizes information about the clusters of companies as concerns the other organizational and managerial factors, allowing a deeper understanding of their business model. Also on these factors, it is possible to appreciate a growing complexity and richness passing organizational factors by firms named C to the ID, type I and II.

Category	C	ID+ type I	ID+ type II
NPD firm organisation	Hierarchical; internal inter-functional teams	Hierarchical **or** flat; internal inter-functional teams	Flat, inter-functional teams with client
Services offered	Marketing, Industrial design, mechanical design, electronics – software, engineering	Marketing, Industrial design, mechanical design, partially engineering	Partially marketing, Industrial design, mechanical design, partially engineering
Competences	Only internal; external rapid prototyping	Internal competencies **or** effective external network (in both cases external rapid prototyping)	Only internal
Project average duration	One year	2 – 10 months	2 – 8 months
Competitors	International	None **or** other NPD firms	Local specialised firms
Clients business sectors	Sport and medical equipments, household appliances, machinery, illumination, aerospace	Sport and medical equipments, cosmetics, household appliances, machinery, aerospace, automotive, yachts, motorcycles, telecommunication, toys	sport and medical equipments, cosmetics, household appliances, machinery, aerospace, automotive, telecommunication, motorcycles, toys
Location of the clients	International	International **or** local	Local
Clients searching	Personal acquaintances, word of mouth, marketing/trade dedicated employees, Web site, presence at professional fairs	Personal acquaintances, word of mouth, marketing/trade dedicated employees, Web site	Personal acquaintances, word of mouth, professional fairs
Well-known brand	No. Well – know only to professionals	Only in few cases	No

Table 3. Main distinctive features and firms clusters

Companies belonging to cluster C are normally big firms with an articulated and complex organizational structure, whilst the others are usually smaller and, above all, showing an informal organization. Consequently, such "complete" firms can boast competences that are organized in functional units able to work together by sharing expertise along the whole NPD process. Thus, the range of services offered is very broad and the level of internal expertise is wide and deep, with advanced skills also in the electronic and software fields. The scope of these enterprises is generally international (in terms of both customers and competitors' nature) and that is what differentiates them from others in particular, although the clients' business sectors are usually very similar. In short, the complexity is really more and this is also evidenced by the longer duration of the undertaken projects. Moreover, although the mode of finding customers seem to be rather informal or quite traditional for all types of firms (i.e. personal acquaintances, word of mouth, marketing/trade dedicated employees) more sophisticated means of communication seems to be spreading (i.e. through Web site) in the "complete" companies. Finally, it is nevertheless interesting to note that the reputation of all the studied companies is strictly limited to professionals and there are very few cases where the brand is known and therefore perhaps rewarding (i.e. Pininfarina Extra and Zagato). However, this confirms the importance of enriching the empirical evidence on this type of service companies.
In the end, as briefly synthesised in table 4, many analogies have been found among NPD service companies. At the same time, several significant differences have emerged, as discussed above.

ANALOGIES		DIFFERENCES	
Human resources management	Interaction with client	Firm organisation and services offered	Project average duration
Knowledge management	Commercialisation phase	Competencies and collaborations	Inter-functional teams
Firm structure	Intellectual property management	Clients' business sectors	
Sale of NPD services and external communication	Performance measurement system	Clients searching	
Customer relationship management	Project management	Competitors	
Pricing		Location of the clients	

Table 4. Analogies and differences among the studied NPD firms

5. Conclusions and future research

In coherence with the main objective stated in the introduction, the paper illustrates a set of Italian firms offering services for New Product Development, from a firm-perspective. The business model adopted by these companies is described in terms of organisation, internationalisation and commercialisation strategy, organisation and management of the

NPD projects. This result represents a step further in the literature that tries to study the emerging "market of technology" from a firm-perspective.
The analysis of the firms' business models can be useful for firms operating in the NPD services market, since it increases the knowledge about this sector and identifies the main features and capabilities that should be developed to pursue such business models.
Moreover, it can be relevant for firms searching for services to support their internal NPD processes. As a matter of fact, they usually completely ignore which kind of services they may find available in the NPD service market and how these NPD companies actually carry out their business. In other words, a better knowledge of the NPD service companies may facilitate an efficient and effective relationship with their potential clients. The identification of clusters is particularly relevant from this point of view: each cluster clearly identifies a specific set of services offered and a definite approach to the management of the business, so that each company searching for external support can select the "set" of potential suppliers most adequate to its specific needs and characteristics.
In terms of future research, the aim is threefold in order to deepen the study of:

- NPD services dedicated to specific sectors of activity, in order to better understand the impact of industry-specific characteristics on the development of services for NPD;
- NPD companies in different countries than Italy;
- Firms covering one phase of the new product development (design companies, engineering companies, prototyping companies etc.).

These studies would offer the opportunity to conduct cross-industry and cross-country comparisons and to deepen the analysis of the main organisational and managerial features highlighted in this paper.

6. References

Abetti, P.A. (1989). Technology: a key strategic resource. *Management Review*, Vol. 78, No. 2, pp. 37-41.

Arora, A.; Fosfuri, A. & Gambardella A. (2001). *Markets for Technology: the Economics of Innovation and Corporate Strategy*. MIT Press, Cambridge.

Chiaroni, D.; Chiesa, V.; De Massis, A. & Frattini F. (2008). The knowledge-bridging role of Technical and Scientific Services in knowledge-intensive industries. *International Journal of Technology Management*, Vol. 41, No. 3-4, pp. 249-272.

Chiesa, V. & Manzini, R. (2001). Innovation and the growing market for technical and scientific services. *Proceedings of the Workshop on Management and Innovation Services*, Maastricht, 5-6 April.

Chiesa, V.; Frattini, F. & Manzini R. (2008). Managing and organising technical and scientific service firms: a taxonomy and an empirical study. *International Journal of Services Technology and Management*, Vol. 10, No. 2/3/4, pp. 211-234.

Chiesa, V.; Manzini, R. & Pizzurno, E. (2004). The externalisation of R&D activities and the growing market of product development services. *R&D Management*, Vol. 34, No. 1, pp. 65-75.

Cooper, R.G. (2008). Perspective: The stage-gates idea-to-launch process—Update, what's new, and nexGen Systems. *Journal of Product Innovation Management*, Vol. 25, No. 3, pp. 213–232.

Debackere, K. (1999). *Technologies to develop technology*, Nijgmen Business School.

Eisenhardt, K.M. (1989). Building theories from case study research. *Academy of management review*, Vol. 14, No. 4, pp. 532-550.

Emes, M.; Hughes, I. & Smith, A. (2005). Internal invention, external development, *Proceedings of the R&D Management Conference* 2005, Pisa July 5-8.

Ginsberg, A. & Abrahamson, E. (1991). Champions of change and strategic shifts: the role of internal and external change advocates. *Journal of Management Studies*, Vol. 28, No. 2, pp. 173-190.

Hargadon, A. & Sutton, R. (1997). Technology brokering and innovation in a product development firm. *Administrative Science Quarterly* Dec. 1997 Vol. 42, No. 4, pp. 716-750.

Horgos, D. & Koch, A. (2008). The internal differentiation of the KIBS sector: empirical evidence from cluster analysis. *International Journal of Services Technology and Management*, Vol. 10, No. 2/3/4, pp. 190-210.

Howells, J. (1999). Research and technology outsourcing. *Technology Analysis and Strategic Management*, Vol. 11, No. 1, pp. 17-29.

Jones, O. & Stevens, G. (1999). Evaluating failure in the innovation process: the micropolitics of new product development. *R&D Management,* Vol. 29, No. 2, pp. 167-176.

Jurgens, U. (2000). *New product development and production networks*. Springer, Berlin.

Kelley, T. (2001). *The art of innovation - lessons in creativity from IDEO, American's leading design firm*. Doubleday, New York.

Knoben, J. & Oerlemans, L.A.G. (2006). Proximity and inter-organizational collaboration: A literature review. *International Journal of Management Reviews*. Vol. 8, No. 2, pp. 71-89.

Kotler, P. (1997). *Marketing management analysis: planning, implementation and control*. Prentice Hall, New Jersey.

Larsen, J.N. (2000). Supplier-User Interaction in Knowledge-Intensive Businee Services: Types of Expertise and Modes of Organization, in Boden, M., Miles, I., *Services and the Knowledge-Based Economy*. Continuum, London.

Lazzarotti, V. & Manzini, R. (2009). Different modes of open innovation: a theoretical framework and an empirical study. *International Journal of Innovation Management,* Vol. 13, No. 4, pp. 615-636.

MacPherson, A. (1997a). The contribution of external services inputs to the product development efforts of small manufacturing firms. *R&D Management*, Vol. 27, No. 2, pp. 127-144.

MacPherson, A. (1997b). The role of external technical support in the innovation performance of scientific instruments firms: empirical evidence from New York State. *Technovation*, Vol. 17, No. 3, pp. 141-150.

Miles, I. (2000). Services innovation: coming of age in the knowledge-based economy. *International Journal of Innovation Management,* Vol. 4, No. 4, pp. 371-389.

Muller, E. & Zenker, A. (2001). Business services as actors of knowledge transformation: the role of KIBS in regional and national innovation systems. *Research Policy*, Vol. 30, No. 9, pp. 1501-1516.

Quinn, J.B. (1999). Strategic outsourcing: leveraging knowledge capabilities. *Sloan Management Review*, Vol. 40, No. 4, pp. 9-21.

Quinn, J.B. (2000). Outsourcing innovation: the new engine of growth. *Sloan Management Review*, Vol. 41, No. 4, pp. 13-28.

Strambach, S. (2008). Knowledge-Intensive Business Services KIBS as drivers of multilevel knowledge dynamics. *International Journal of Services Technology and Management*, Vol. 10, No. 2/3/4, pp. 152 -174.

Trott, P. (2008). *Innovation management and new product development*. Prentice Hall, New Jersey.

Verganti, R. (1997). *R&D Management*. Blackwell Publishers, Oxford.

Veugelers, R. & Cassiman, B. (1999). Make and Buy in innovation strategies: evidence from Belgian manufacturing firms, *Research Policy*, Vol. 28, No. 1, pp. 63-80.

Windrum, P. & Tomlison, M. (1999). Knowledge-intensive Services and international competitiveness: a four-country comparison. *Technology Analysis and Strategic Management*, Vol. 11, No. 3, pp. 391-405.

Yin, R. K. (2003). *Case study research, design and methods*. 3rd ed. Sage Publications, Thousand Oaks.

Zenker, A. & Doloreux, D. (2008). KIBS, perceptions and innovation patterns. *International Journal of Services Technology and Management*, Vol. 10, No. 2/3/4, pp. 337-342.

Zhao, Y. & Calantone, R. J. (2003). The trend towards outsourcing in new product development: case studies in six firms. *International Journal of Innovation Management* Vol. 7, No. 1, pp. 51-66.

9

Process Capability and Six Sigma Methodology Including Fuzzy and Lean Approaches

Özlem Şenvar and Hakan Tozan
Marmara University, Turkish Naval Academy
Turkey

1. Introduction

Process capability analysis (PCA) and Six Sigma methodology occupy important places in quality and process improvement initiatives. As a fundamental technique in any production, quality and process improvement efforts, PCA is used to improve processes, products or services to achieve higher levels of customer satisfaction. In order to measure process capability numerically, process capability indices (PCIs) have been developed. Six Sigma is widely recognized as a systematic methodology that employs statistical and non-statistical tools and techniques for continuous quality and process improvement and for managing operational excellence because it challenges to maximize an organization's return on investment (ROI) through the elimination of nonconforming units or mistakes in the processes (Antony et al., 2005). The application of Six Sigma methodology provides reduction in variance and augmentation in the process capability, which is defined as the proportion of actual process spread to the allowable process spread that is measured by six process standard deviation units. Similar to Six Sigma methodology, in a process capability study, the number of standard deviations between the process mean and the nearest specification limits is given in sigma units. The sigma quality level of a process can be used to express its capability that means how well it performs with respect to specifications.

After Zadeh (1965) introduced the Fuzzy Logic (FL) to the scientific world, this new phenomenon rapidly became an essential systematic used in nearly every field of science. Due to its capability of data processing using partial set membership functions, an enormous literature about FL is developed with full of its applications. In addition, the ability of donating intermediate values between the expressions mathematically turns FL into a strong device for impersonating the ambiguous and uncertain linguistic knowledge (Ross, 2004). But although studies about FL are extremely wide, its application to quality control and especially to PCA is relatively narrow.

The aim of this chapter is to carry out a literature review of PCA, fuzzy PCA, PCIs, to make comparisons between PCIs, to introduce ppm and Taguchi Loss Function, to discuss the effects of estimation on PCIs as well as to provide general discussion about sample size determination for estimating PCIs. Another objective of this chapter is to provide the investigation of the relationship between Process Capability and Six Sigma along with the examination of Six Sigma methodology, and a relatively new approach called Lean Six Sigma methodology, and to identify the key factors that influence the success of Six Sigma project implementation for improving overall management process.

2. Process capability

2.1 Process

Process is defined as a combination of materials, methods, equipments and people engaged in producing a measurable output. As a matter of fact, all processes have inherent statistical variability, which can be identified, evaluated and reduced by statistical methods.

The source and amount of variability should always be considered by organizations. In order to satisfy customer requirements, organizations must improve the quality by reducing variance in production processes. The less variation the system has, the better quality it provides. Thereby, the variability of critical-to-quality characteristics (CTQs) is a measure of the uniformity of outputs. When the variation is large, the numbers of products that are nonconforming are large. Nonconforming (NC) is the failure of meeting specification limits whereas specifications are the desired measurements for a quality characteristic.

2.2 Process capability

In particular, process capability deals with the uniformity of the process. Variability of CTQs in the process is a measure of the uniformity of outputs. Here, variability can be thought in two ways: one is inherent variability in a CTQ at a specified time, and the other is variability in a CTQ over time. It should be considered that process capability study frequently measures functional parameters or CTQs on the product. It does not measure the process itself (Montgomery, 2009). Process capability compares inherent variability in a process with the specifications that are determined according to the customer requirements. In other words, process capability is the proportion of actual process spread to the allowable process spread, which is measured by six process standard deviation units. Process capability compares the output of a process that is an in-control state to the specification limits by using PCIs. To sum up, a capable process is the one where almost all the measurements fall inside the specification limits and process capability study can be conducted to indicate the extent to which the process can meet these specifications.

In a true process capability study, when there is direct observation of the process, inferences can be made about the stability of the process over time by directly controlling or monitoring data collection activity and understanding the time sequence of the data. However, when there is no direct observation of the process, only sample units of product are known, in this case, the study is called product characterization. In a product characterization study, distribution of the product quality characteristic or the fraction that conforms to specifications, which is referred to as process yield, can only be estimated, notably information about stability or dynamic behavior of the process cannot be given (Montgomery, 2009).

2.3 Process Capability Analysis (PCA)

PCA involves statistical techniques, which are useful throughout the product cycle. Generally, PCA is used in development activities prior to manufacturing process, in quantification of process variability, in analysis of this variability relative to specifications and in elimination or reduction of the process variability (Montgomery, 2009).

As a fundamental technique in any production, quality and process improvement efforts, PCA is used to improve processes, products or services to achieve higher levels of customer satisfaction. PCA has become widely adopted as the measure of performance to evaluate the ability of a process to satisfy customer requirements in terms of specification limits (English

& Taylor, 1993). The output of a process is expected to meet specifications, which can be determined according to the customer requirements. PCA is a prominent technique that is used to determine how well a process meets to these specification limits. PCA is based on a sample of data taken from a process and often produces: an estimate of the dpmo (defects per million opportunities), one or more capability indices, an estimate of the sigma quality level at which the process operates. The sigma quality level of a process can be used to express its capability that means how well it performs with respect to specifications.

As a measure of process capability, it is customary to take six sigma spread in the distribution of product quality characteristic. For a process whose quality characteristic has a normal distribution with process mean μ and process standard deviation σ; the lower natural tolerance limit of the process is LNTL = μ – 3σ, and the upper natural tolerance limit of the process is UNTL = μ + 3σ. It should be considered that natural tolerance limits include 99.73% of the variable and 0.27% of the process output falls outside the natural tolerance limits.

PCA is often used to estimate the process capability. The estimate of process capability can be in the form of a distribution that has parameters of shape, center (mean) and spread (standard deviation). In this case, PCA can be performed without regard to specifications of the quality characteristic. Here, process capability can be expressed as a percentage outside of specifications (Montgomery, 2009). For PCA, the following techniques can be used:

- **Histograms:** In statistics, histograms are defined as graphical displays of frequencies. In the quality applications, histograms are well-known as one of the seven basic tools of quality control. Histograms are very useful in estimating process capability and for visualizing process performance. Hence, histograms can be used to determine the reason for poor process performance, instantaneously. As quality characteristics are often assumed to have normal distribution, histogram along with the sample mean and sample standard deviation can provide information about process capability as it is possible to estimate the process capability independent of the specifications. Here, normality assumption can be investigated by looking at the shape of the histogram. If the histogram is fairly skewed, then the normality assumption might be a concern and estimate of the process capability is unlikely to be correct. On the other hand, there are some drawbacks of using histograms. Fundamentally, it is necessary to divide the range of a variable into classes. Also, histograms cannot be used for small samples, for this reason, at least hundred observations are needed. Essentially, in order to have reliable estimate of process capability, these observations must be moderately stable (Montgomery, 2009).
- **Probability Plots:** Probability plots are very useful in estimating the process capability. Also, probability plots can be used to determine distribution's parameters, which are shape, center and spread. Furthermore, it is unnecessary to divide the range of a variable into class intervals. Probability plots can be used for moderately small samples, as well. However, if the data of quality characteristic do not come from the assumed distribution, inferences about process capability may be seriously in error. That can be shown as drawback of probability plots as they are not objective procedures. Practically, normal probability plots are very useful in process capability studies (Montgomery, 2009). Here, the fat pencil test is preferred to be used for testing the adequacy of the normality assumption. The fat pencil test is performed like that: when the data are plotted against a theoretical normal distribution, the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

- **Design of Experiments (DOE):** DOE is very useful for identifying critical parameters associated with a process and determining optimal settings for these process parameters for enhanced capability and performance of the process. In other words, DOE is a systematic approach that is carried out to vary the input controllable variables in the process and analyze the effects of these process variables on the output, which is referred to as response in the DOE terminology. DOE is used to discover which set of process variables is influential on the output, and at what levels these variables should be held to optimize process performance. One of the major uses of DOE is discriminating and estimating the sources of variability in a process (Montgomery, 2009). Literally, DOE has been widely accepted in manufacturing processes and is useful in more general problems rather than merely estimating the process capability.
- **Control Charts:** Control charts are very useful for establishing a baseline of the process capability or process performance. Control charts can be used as monitoring devices to show effects of changes in the process on process performance. Basically, control charts can determine whether a manufacturing or business process is in a state of statistical control or not. They show systematic patterns in process output, as well. In particular, before using PCIs, there is a need for establishment of a state of statistical control. That is, if a control chart indicates that the process is currently under control, then it can be used with confidence to predict the future performance of the process. In the contrary, if a control chart indicates that the process being monitored is not in control, the pattern it reveals can help to determine the source of variation to be eliminated in order to bring the process back into control. Concisely, the control chart allows significant change to be differentiated from the natural variability of the process. This is shown to be the key for effective process control and improvement. Control charts are effective in displaying potential capability of the process by performing the issue of statistical control, for this reason, they should be regarded as the primary technique of PCA. In PCA, both variables and attributes control charts can be used (Montgomery, 2009).

2.4 Process Capability Indices (PCIs)

In the literature, process capability indices (PCIs) are also called process capability ratios (PCRs). PCIs are used as tools for characterizing the process quality. In order to measure the process capability numerically, PCIs have been developed. PCIs use process specifications as well as process variability, in this regard, the use of PCIs is important as they are statistical indicators of the process capability. PCIs are also defined as the quantitative indicators that compare the behavior of process or product characteristic to the specifications. In other words, PCIs are used to determine how well the process performs with respect to specifications and they express the ability of the process to meet these specifications, as a unique value quantitatively.

There are several statistics that can be used to measure the capability of a process. Frequently used measures of performance are the PCIs, which relate the natural tolerance limits of a process to the specification limits (English & Taylor, 1993). In practice, Cp, Cpk (Cpl, Cpu), Cpm are some of the widely used PCIs. In next sections, process capability indices: Cp, Cpk (Cpl, Cpu), Cpm, Cpmk will be explained.

2.4.1 Process Capability index Cp

In the literature, Cp index is also called process potential index, or process capability ratio, or inherent capability index, and two-sided PCI for two-sided specifications, that is, process

is having both lower and upper specification limits. Cp is frequently used in industrial environment in order to express process capability in a simple quantitative way. When the parameters are known, that is, in that case, when process standard deviation σ is known, PCI Cp is computed as follows:

$$Cp = \frac{USL\text{-}LSL}{6\sigma} \tag{1}$$

where LSL and USL are lower and upper specification limits, respectively. The percentage of the speciation band used up by the process can be calculated in the following way:

$$P=\left(\frac{1}{Cp}\right)*100 \tag{2}$$

In practice, it is often impossible to know parameters. Generally, it is suitable to use sample standard deviation s to estimate process standard deviation σ. Thus, when the parameters are unknown, that is, in that case, when process standard deviation σ is unknown, by replacing sample standard deviation s to estimate process standard deviation σ, the formula used for estimating Cp is given below:

$$\widehat{Cp} = \frac{USL\text{-}LSL}{6s} \tag{3}$$

where LSL and USL are lower and upper specification limits, respectively.
A Cp value less than 1 indicates that the process variation exceeds the specifications and a significant number of defects are made. A Cp value equal to 1 indicates that the process is exactly meeting the specifications. At least 3% defects would be made. However, if the process is not centered on the target value (off-center), more defects are expected to be made. A Cp value greater than 1 indicates that the process variation is less than the specifications. However, if the process is not centered on the target value (off-center), more defects are expected to be made. A Cp value greater than 1.67 indicates that the process is highly capable.

2.4.2 Process Capability index Cpk

In the literature, for one-sided specifications, Cpk is defined as one-sided PCI for specification limit nearest to the process mean. When the parameters are known, that is, in that case, when process mean μ and process standard deviation σ are known, PCI Cpk is computed as follows:

$$Cpk = \frac{1}{3\sigma}\min(USL\text{-}\mu, \mu - LSL)=\min(Cpu, Cpl) \tag{4}$$

where LSL and USL are lower and upper specification limits, respectively. In practice, it is often impossible to know parameters. Generally, it is suitable to use sample mean $\bar{x}$ to estimate process mean μ and sample standard deviation s to estimate process standard deviation σ. When the parameters are unknown, that is, in that case, when process mean μ and process standard deviation σ are unknown, by replacing sample mean $\bar{x}$ and sample standard deviation s to estimate process mean μ and process standard deviation σ, respectively, the formula used for estimating Cpk is given below:

$$\widehat{Cpk}=\frac{1}{3s}*\min(USL-\bar{x},\ \bar{x}-LSL)=\min(Cpu,Cpl) \tag{5}$$

where LSL and USL are lower and upper specification limits, respectively.
Montgomery (2009) defined Cp as the measurement of the potential capability in the process. As a matter of fact, Cp does not consider where the process mean is located relative to the specification limits. Cp only measures the spread of the specifications relative to the six sigma spread in the process. Cp does not deal with the case of a process with mean μ that is not centered between the specification limits. On the other hand, he defined Cpk as the measurement of the actual capability in the process. Cpk takes process centering into account. In other words, Cpk deals with the case of a process with mean μ that is not centered between the specification limits. The magnitude of Cpk relative to Cp is the direct measure of how off-center the process is operating. Montgomery (2009) examined several cases, which can explain the relationship between Cp and Cpk, are given below:

- If Cp=Cpk, the process is centered at the midpoint of the specification limits.
- If Cpk<Cp, the process if off-centered. This can be accepted as lower capability than the case that the process is centered. The reason is that it is not operating at the midpoint of the interval between the specification limits.
- If Cpk=0, the process mean is exactly equal to one of the specification limits.
- If Cpk<0, the process mean lies outside the specification limits, that is for μ>USL or μ<LSL, Cpk<0.
- If Cpk<-1, the entire process lies outside the specification limits. It should be noted that some authors define Cpk to be nonnegative so that values less than zero are defined as zero.

1<Cpk<1.33 means that the process is barely capable. Automotive industry uses Cpk=1.33 as a benchmark in accessing the capability of a process (AIAG, 2002).

2.4.3 Process Capability index Cpm

In the literature, Cpm is referred to as Taguchi index. Simply, Cpm is defined as the ability of the process to be clustered around the target or nominal value, which is the measurement that meets to exact desired value for the quality characteristic. Actually, Cpm was developed because Cpk is observed to be inadequate measure of process centering although Cpk was developed to deal with the case of a process with mean μ that is not centered between the specification limits wheras Cp is inadequate in process centering. As a matter of fact, when μ is in the interval of the specification limits, LSL and USL, Cpk depends inversely on process standard deviation σ and becomes large as process standard deviation σ gets closer to zero. Keeping these features in mind, it is possible to say that Cpk is not convenient as a measure of centering. This means a large value of Cpk does not actually give any information about the location of the mean in the interval of the specification limits, LSL and USL. In that case, process capability index Cpm, which is a better indicator of process centering, would be much more convenient (Montgomery, 2009). Consequently, the PCI Cpm is intended to account for variability from the process mean and deviation from the target value T and Cpm is shown to be useful in process centering. When the parameters are known, that is, in that case, parameters of process mean μ and process standard deviation σ are known, PCI Cpm is computed as follows:

$$Cpm=\frac{USL-LSL}{6\tau} \tag{6}$$

where τ is the square root of expected squared deviation from target T. The target value T, which is the measurement that meets to exact desired value for the quality characteristic, is known to be the midpoint of the specification interval. Target T is evaluated as follows:

$$T=\frac{1}{2}(LSL+USL) \tag{7}$$

The formula for process variation around desired process target is given below:

$$\tau^2=E\left[(x-T)^2\right]=E\left[(x-\mu)^2\right]+(\mu-T)^2=\sigma^2+(\mu-T)^2 \tag{8}$$

Computation of Cpm can also be performed with the following way:

$$Cpm=\frac{USL-LSL}{6\sqrt{\sigma^2+(\mu-T)^2}}=\frac{Cp}{\sqrt{1+\left(\frac{\mu-T}{\sigma}\right)^2}} \tag{9}$$

Cpm approaches zero asymtotically as $|\mu- T|\rightarrow\infty$. When the parameters are unknown, that is, in that case, when process mean μ and process standard deviation σ are unknown, by replacing sample mean $\bar{x}$ and sample standard deviation s to estimate process mean μ and process standard deviation σ, respectively, the formulas used for estimating PCI Cpm is given below:

$$\widehat{Cpm}=\frac{\widehat{Cp}}{\sqrt{1+V^2}} \tag{10}$$

where; $V=\frac{\bar{x}-T}{s}$.

2.4.4 Process Capability index Cpkm

The motivation of Cpkm is increased sensitivity to departures of the process mean μ from the desired target value T. Cpkm is known as a third generation PCI, since it is derived from the second generation PCIs Cpk and Cpm, in the same way that the PCIs, Cpk and Cpm are derived from the first generation PCI Cp. Computation of Cpkm is as follows:

$$Cpkm=\frac{Cpk}{\sqrt{1+\left(\frac{\mu-T}{\sigma}\right)^2}} \tag{11}$$

At the end of this section, it has to be emphasized that PCIs can measure expected future performance. Industrial use of PCIs concentrates on evaluating and interpreting the point estimates of the desired quantities of PCIs, which are utilized to measure the ability of a process to meet the specification limits. It must be noted that point estimates of PCIs are simply point estimates and they are subject to statistical fluctuation. In other words, since point estimates of PCIs are subject to variability, alternatively, researchers recommend practitioners to use confidence intervals for estimating PCIs. There is a recent focus on

hypothesis testing and confidence intervals on PCIs that are used as the basis for establishing the process capability (English & Taylor, 1993). For details about hypothesis testing and confidence intervals on PCIs, interested readers are referred to Montgomery, 2009.

2.5 Comparisons between PCIs

In the review paper of Kotz and Johnson (2002), Cp is ascribed to Juran , Cpk to Kane, and Cpm for the most part to Hsiang and Taguchi. Kotz and Johnson emphasized that it is necessary to distinguish the features of PCIs and the features of their estimators. Apart from this, the relationship between these PCIs are defined as; "**$Cp \geq Cpk$ and $Cp \geq Cpm$**". Also, researchers realize that Cpk and Cpm coincide with Cp when $\mu=T$ and decrease as μ moves away from target T, wheras $Cpk<0$ for $\mu<LSL$ or $\mu>USL$.

Spiring et al. (2002) highlighted that both Cp and Cpk are related to expected proportion of nonconforming items or defects. In other words, Cp and Cpk are related to marginal expected value of ppm (parts per million). On the other hand, Cpm does not arise from examining the number of nonconforming product in the process. Therefore, Cpm is unreliable if the expected proportion of nonconforming is regarded as the most important feature. Unlike the other PCIs, Cpm is not distributionally sensitive.

In industrial practice, it should be noticed that the motivation of Cp, Cpl, Cpu, Cpk are the most extensively used PCIs, while Cpm is seldomly being used. According to Bothe (2002), Cpk seems to have the greatest degree of acceptability among the PCIs. It is important to emphasize that Cpk is not suitable for product features with asymmetric tolerances. Even all the assumptions are satisfied, a higher Cpk does not represent a higher level of quality for customers. On the other hand, Cpm is related to Taguchi quadratic loss function because Cpm is defined as the ability of the process to be clustered around the target. Furthermore, Cp, Cpl, Cpu, Cpk are interpreted as the measure of nonconforming. Any change in the magnitude of these indices, under the constraint of holding customer requirements constant, is due to changes in the distance between the specification limits and the process mean. Cpk does not in itself say anything about distance between μ and T and it only measures the process yield (Spiring et al., 2002).

2.6 Taguchi loss function

PCA examines the ability of a process to satisfy customers in terms of specification limits. However, sometimes, it can be more suitable to investigate the costs associated with process variation. For this purpose, Taguchi quadratic loss function can be used in order to examine the costs. In other words, Taguchi loss function is generally preferred to be used in modeling the expected costs. The basis of the Taguchi quadratic loss function is incurred when the quality characteristics of a product deviate from the target value. Taguchi loss function is shown below:

$$L=k*(Y-T)^2 \tag{12}$$

where L symbolizes loss function; k is constant; Y is the observed value of the quality characteristic, and T is the target value of quality characteristic.

English and Taylor (1993) report that the target value for the quality parameter as a design variable can be adjusted easily. Process design engineers can alter the idea of utilization of the specification and utilize more optimal target values supported by known process behavior.

Taguchi used the quadratic loss function for motivating the idea that a product imparts no loss only if it is produced at its target. He maintained that even small deviations from the target result in a loss of quality. Taguchi's philosophy highlights the need to have low variability around the target. As a result of this, the most capable process produces its product at the target.
Taguchi identified that when "nominal is the best" is assumed, the expected value of loss minimized regardless of the distribution. In that case, the target value is adjusted to be equal to the expected value of the underlying process. It should be noticed that, by stabilizing the process and reducing the variation will reduce the cost of the Taguchi loss function. Taguchi loss function strategy emphasizes reducing variability and striving for a process mean that equates to the nominal specification.
In fact, PCIs are based on expected loss. Quality improvement efforts deal with reducing variances and discriminating them as much as possible. For this purpose, there is an increasing importance of clustering around the target rather than conforming to the specification limits. This makes Taguchi loss function to be an alternative to PCIs. Production costs or losses can provide opportunities to access, monitor and compare process capability (Spring et al., 2002). For more details on the topic, interested readers are referred to English & Taylor (1993).

2.7 The effects of estimation on PCIs

Theoretically, PCIs can be computed when the values of the process parameters (process mean μ and process standard deviation σ) are known. However, in practice, these process parameters required for computing PCIs are almost always unknown, for this reason, estimation is used in evaluating the process capability. In order to evaluate the process capability, these unknown process parameters have to be estimated from a sample of observations from the process but this is known to have effects on the estimated values of the PCIs.
Often, after PCIs are computed, these indices are converted into measures such as ppm (parts per million) defective items, also known as nonconforming items. The estimate of expected proportion of the nonconforming items, which is symbolized as p, can be obtained from the tails of the corresponding distribution exceeding the specification limits.
It should be noticed that estimated values are subject to variation and these values are generally different from the actual values. In other words, the estimate is subject to error in estimation as it depends on sample statistics (Montgomery, 2009). As a matter of fact, quality of the estimation is an important issue for the reliability of the calculated statistics such as Cp and ppm. Therefore, quality of the estimation should be considered for the reliability of the estimated PCIs and ppm statistics.
PCIs can only be used when the process is in the state of statistical control. Similarly, in order to have a reliable estimate of process capability, the process should be stable or be in statistical control. Stability or statistical control of the process is really necessary for correct interpretations of the PCIs. If the process is not in statistical control, then its parameters would be unstable. As a result, estimates of these parameters would be uncertain. Thus, predictive aspects of the PCIs regarding process ppm performance are not valid at all (Montgomery, 2009). Unless the process is stable (in control), no index is going to carry useful predictive information about process capability. Ramirez & Runger (2006) pointed out that the fundamental concepts of "in-control" and "capable" are the pillars of the process capability control. According to Kotz & Johnson (2002), the assumption of attaining

state of statistical control of the process is required in order to detect the irregular changes in quality level. Regardless of how robust an estimator may be, if its associated parameter is not stable, then any robustness claims carry little meaning (Spiring et al., 2002). PCA can determine an out-of-control process. In such cases, it is not safe to estimate process capability. When the process is out-of-control at the beginning of the PCA, it is possible to bring the process into an in-control state by eliminating assignable causes. Keeping all these features in mind, it is to be emphasized that for predictive and stable processes, the PCIs can measure the expected future performance.

Apart from the stability assumption, the quality characteristic is assumed to have a normal distribution with a constant mean and variance. Checking the normality assumption of the data is essential for meaningful interpretations of the PCIs. Lack of normality assumption may provide misleading interpretations of the results. When the process output is normally distributed, there is a certain relationship between a given values of Cp and the expected proportion of nonconforming items produced by the process monitored. As a matter of fact, normality assumption is very important because interpretation of process capability and PCIs, especially Cp and Cpk, depend on normal distribution of process output. According to English & Taylor (1993), Cp, Cpk, Cpm statistics assume that the process measurements are independently and identically distributed normal, that is iid $N(\mu, \sigma^2)$. Thus, meaningful interpretation of the indices of Cp, Cpk, Cpm is based on the normal distribution assumption. If the underlying distribution is non-normal, then expected process fallout attributed to a specific value of Cp, Cpk, Cpm may be seriously in error.

In essence, when the PCIs are estimated appropriately, they provide important information about how the current process meets customer requirements or the specifications. In contrast to this, incorrect application or interpretation of the PCIs causes unreliable results, which can lead incorrect decision making, waste of resources, money, time, and etc.

It has to be emphasized one more time that PCIs can measure expected future performance. Industrial use of PCIs concentrates on evaluating and interpreting the point estimates of the desired quantities of PCIs, which are utilized to measure the ability of a process to meet the specification limits of the customer requirements. It must be noted that point estimates of PCIs are simply point estimates and they are subject to statistical fluctuation. In other words, since point estimates of PCIs are subject to variability, alternatively, researchers recommend practitioners to use confidence intervals for estimating PCIs. There is a recent focus on hypothesis testing and confidence intervals on PCIs that are used as the basis for establishing the process capability (English & Taylor, 1993).

2.8 General Discussion about Sample Size Determination for Estimating PCIs

The estimate of PCI is always subject to error since it depends on sample statistics. English & Taylor (1993) mentioned that estimating PCI from sample data can cause large errors. As a result of this, the estimate may not be reliable, at all. For convenience, utilization of confidence intervals for estimating PCI should be considered. As point estimators for PCIs are subject to errors, point estimates of PCIs may not be satisfactory, if they are computed from small samples. In other words, point estimates are useless if they are computed from small samples. Similarly, confidence intervals on PCIs with small samples will always be wide, which are not preferable.

Kotz & Johnson (2002) mentioned that there are recent investigations about asymptotic properties of estimators that indicate the importance of determining sample sizes n for which asymptotic results are adequate.

As a matter of fact, point estimates are subject to variability and would change over time, even process remains stable. This can be an advantage for computation of the confidence limits for process control. It should be emphasized that checking normality assumption of the data is necessary for interpretations of PCIs and for the validity of confidence limits.
Considerably, t-distribution with as many as 30 degrees of freedom is symmetric and visually indistinguishable from the normal distribution. Nevertheless, symmetry in the distribution of the process output alone is not sufficient to ensure PCI would provide a reliable estimate of processs ppm. Thereby, the longer and heavier tails of t-distribution is making meaningful difference, when ppm is being estimated (Montgomery, 2009).
Notice that, when the process output is normally distributed, there is a certain relationship between a given values of Cp and the expected proportion of nonconforming items produced by the process monitored. Quality of the estimated Cp and ppm values depends on the sample size used in the estimation. As a matter of fact, in practice, the quality of the estimates of PCI, such as Cp, can be changed according to the sample size (Deleryd, 1999). Therefore, large samples are required to be used to obtain reliable estimates.

3. Fuzzy Process Capabilities Analysis (Fuzzy PCA)

Since pioneer work of Zadeh (1965), fuzzy logic (FL) has been successfully applied to many fields of science and engineering. Studies in quality and PCAs domain have also effected from researches that involves the application of FL; especially from ones which have been applied to statistical methods such as quality control (Wang et al., 1990; Faraz et al., 2006; Gulbay et al., 2006 and 2007). Thought after 2000, studies about PCAs and its integral part PCIs from the FL point of view began to grab attention and stepped up; beside of the enormous crisp literature of PCAs, they are relatively in the minority.
The elementary idea of using FL approach for PCAs and PCIs can simply be express as, to overcome infirmity of PCIs arisen from the sharp crisp nature that restricts the flexibility, applicability and sensitivity which; both, individually and together directly affect the performance of PCAs. In this section, after a shallow mention to this brilliant subject, a brief summary on the studies about fuzzy PCA and PCIs will be given.

3.1 Fuzzy logic

FL can simply be defined as "a form of mathematical logic in which truth can assume a continuum of values between 0 and 1" (http://wordnetweb.princeton.edu/, 2009). On the contrary to many cases that involves human judgement, crisp (discrete) sets divide the given universe of discourse in to basic two groups; members, which are certainly belonging the set and nonmembers, which certainly are not. This delimitation which arises from their mutually exclusive structure enforces the decision maker to set a clear-cut boundary between the decision variables and alternatives. The basic difference of FL is its capability of data processing using partial set membership functions. This characteristic; including the ability of donating intermediate values between the expressions mathematically, turn FL into a strong device for impersonating the ambiguous and uncertain linguistic knowledge. But the main advantage of fuzzy system theory is its ability "to approximate system behavior where analytic functions or numerical relations do not exist" (Ross, 2004, pg.7). Palit et al. (2005) give a basic definition of FL from mathematical perspective as a nonlinear mapping of an input feature vector into a scalar output. As fuzzy set theory became an important problem modeling and solution technique due to its ability of modeling problems

quantitatively and qualitatively those involve vagueness and imprecision (Kahraman, 2006, pg.2), it has been successfully applied many disciplines such as control systems, decision making, pattern recognition, system modeling and etc. in fields of scientific researches as well as industrial and military applications (Tozan et al., 2008, 2009).
As stated before, differently from the classical sets that can be defined by characteristic functions with crisp boundaries, fuzzy sets can be characterized by membership which provides expressing belongings with gradually smoothed boundaries (Tanaka, 1997). Let A be a set on the on universe X with the objects donated by x in the classical set theory. Then the binary characteristic function of subset A of X is defined as follow;

$$\mu_A(x) : X \to \{0,1\} \tag{13}$$

such that

$$\mu_A(x) = \begin{cases} 1 & x \in X \\ 0 & x \notin X \end{cases} \tag{14}$$

But fuzzy sets the characteristic functions; differently from the crisp sets whose characteristic function is defined binary (i.e., 0 or 1), are defined in the interval of $[0,1]$ (Zadeh, 1965). From this point, fuzzy set $\tilde{A}$ in the universe set X with the objects x and membership function $\mu_{\tilde{A}}$ is defined as follow;

$$\tilde{A} = \left\{ \left(x,\ \mu_{\tilde{A}}(x) \right) \quad \forall\ x \in X \ \right\} \tag{15}$$

where $\mu_{\tilde{A}}(x) : X \to [0,1]$.
If the fuzzy set is discrete then it can be represented as;

$$\tilde{A} = \sum_{k}^{n} \frac{\mu_{\tilde{A}}(x_k)}{x_k} \quad , \quad \forall x_k \in X \ , \ k = 1,2,.....,n \tag{16}$$

And if the fuzzy set is continuous then it can be denoted as;

$$\tilde{A} = \int_{X} \mu_{\tilde{A}}(x_k) \Big/ x_k \quad , \quad \forall x_k \in X \tag{17}$$

The two vital factors for building an appropriate fuzzy set gets through the determination of appropriate universe and membership function that fits the system to be defined. The membership functions are the main fact for fuzzy classification. The highest membership grade value 1 represents full membership while the lowest membership value 0 have the meaning that the defined object have no membership to the defined set. Frequently used membership functions in practice are triangular, trapezoidial, Gaussian, sigmoidal and bell curve (the names are given according to the shapes of the functions). For example, the triangular membership function is specified by parameters $\{a,b,c\}$ as:

$$triangular(x;a,b,c) = \begin{cases} x-a \Big/ b-a & ; \quad a \le x \le b \\ c-x \Big/ c-b & ; \quad b \le x \le c \\ 0 & ; \quad x \ge c \ \text{ or } \ x \le a \end{cases} \tag{18}$$

where $a < b < c$. The width of function changes according to the values of a and b.
Characteristic of fuzzy set plays an important role in fuzzy PCA and PCIs studies. For basic concepts of fuzzy sets and related basic definitions see Bellman et al. (1970), Tanaka (1997 pg.5-44), Klir et al. (1995) and Ross (2004, pg.34-44).

3.2 Fuzzy PCA and PCIs

As stated before, though FLs' broad range of application has also effected studies on PCA and PCIs, the fuzzy based perspective on these areas are relatively new. The first spectacular fuzzy PCI studies; to our knowledge, can be traced back to the fuzzy quality and probability study of Yongting (1996); in which, a fuzzy Cpk was defined to determine the fuzzy quality. Later Lee et al. (1999) declared a fuzzy based model to maximize PCI via determining upper and lower bounds of PCIs using membership functions. In 2001, Lee proposed an estimation approach for fuzzy Cpk using fuzzy observations comprised of fuzzy numbers.
One of the worth mentioning fuzzy based study about process capability evaluation is made by Chen et al. (2003). Chen and his research colleagues proposed a method to interlink PCI with a fuzzy inference system for "bigger-the-best" type evaluation. In the study, the input for the fuzzy inference is the p value is calculated as follow;

$$p - value = p\left\{\hat{C}_{pl}^{\iota} \le V \middle| C_{pl} = C_{\min}\right\} \tag{19}$$

where $\hat{C}_{pl}^{\iota}$ is a uniformly minimum variance unbiased estimator of process capability index (C_{pl}) for a normal distribution, $C_{\min}$ is the minimum process capability required for the "bigger the best" type. The proposed steps of Chen et al.'s study for fuzzy evaluation to specify process capability are:

i. Assigning $C_{\min}$ and lower specification limit;
ii. Deciding the manufacturing allowance and the test of α -risk;
iii. Calculating the mean ($\bar{x}$) and the standard deviation s from the selected n sample data set;
iv. Computing the required parameters for obtaining p value using equation (19) through the cumulative distribution function with a non-central t distribution.
v. Defining the membership functions for input and output and inferring the score value by difuzzification. Here, the membership functions for the input and output variables are defined by linguistic variables. The triangular type membership function is used for the input whereas Gaussian type is used for the output.

After inferring the score with difuzzification, authors used a conscience score concept to represent the grade of process capability. Later in their study about multi process capability plot, they proposed a fuzzy inference system approach which is effective for the assessment of multi process capability (2008).
Parchami and his research colleagues (2005, 2006a, 2006b, 2007, 2008, 2010a, 2010b) have also made several important studies related to fuzzy PCIs and fuzzy quality control. In 2005 they introduced fuzzy PCIs determining the relations governing between PCIs when lower specification limits are fuzzy numbers. Moeti et al. (2006) using lower specification limits as L-R intervals also discussed a generalized version of PCIs introduced in Parchami et al.(2005)'s study. Later for a new PCI; $\tilde{C}_p$, they obtained a fuzzy confidence interval and analyzed process capability based on the introduced fuzzy index. (2006a, 2006b, 2008). In

2007, they proposed a Buckley (2004, 2005a, 2005b) approach based algorithm to determine fuzzy estimates (which contains both point and interval estimates) for PCIs providing more information for the practitioners. Parchami and research colleagues declared that when lower specification limits are fuzzy rather than crisp, traditional PCIs does not suit for process capability measurement and they introduced new indices for the cases in which engineering lower specification limits are fuzzy (2010a, 2010b).

From 2007 till 2010, Kahraman & Kaya have made remarkable studies; which indeed, dynamise the researches on quality control and PCA in fuzzy domain. In 2007, they proposed a methodology for air pollution control by using fuzzy and traditional PCIs (Kaya et al., 2007). Later, they used fuzzy PCIs for controlling pH value of dam's water (Kahraman et al., 2008). They also applied fuzzy PCA to learning processes for faculty courses (Kaya et al., 2008a). Using fuzzy PCIs, they have analyzed the risk assessment of air pollution in largest city of Turkey via measuring air pollutants from different stations deployed in different parts of the city (Kaya et al., 2008b). In 2009, they used fuzzy process accuracy index to evaluate risk assessments of draught effects (Kahraman et al., 2009) and; the same year, with their study on air pollution control, they used fuzzy PCIs in six-sigma approach to prevent air pollution (Kaya, 2009).

In 2010, by defining specification limits and standard deviations with fuzzy numbers, Kaya and Kahraman increased PCIs' flexibility and obtained robust PCIs for a piton manufacturing company (Kaya et al., 2010a). They concluded that, for the cases in which crisp numbers can not be appropriate for defining specification limits, fuzzy numbers can be applied to represent specification limits via which, results gathered from the measurement can be analyzed more flexibly. In the study fuzzy PCIs are obtained using triangular fuzzy numbers (TFN) for defining upper and lower specification limits in addition to fuzzy variances follow.

Let the fuzzy upper specification limit be $\tilde{USL} = (a_1, a_2, a_3)$ and the fuzzy lower specification limit be $\tilde{LSL} = (b_1, b_2, b_3)$. Then, α -cuts for $\tilde{USL}$ and $\tilde{LSL}$ are;

$$\begin{aligned} \tilde{USL}_\alpha &= \left[(a_2 - a_1)\alpha + a_1, (a_2 - a_3)\alpha + a_3\right] \\ \tilde{LSL}_\alpha &= \left[(b_2 - b_1)\alpha + b_1, (b_2 - b_3)\alpha + b_3\right] \end{aligned} \tag{20}$$

Assuming parameters analyzed by PCIs have coloration, fuzzy robust PCIs are derivated in the study with the following formulas.

$$\tilde{C}_{pc} = \left(\frac{\left[(a_2 - a_1) + (b_3 - b_2)\right] x\alpha + (a_1 - b_3)}{6x\sqrt{\Phi_1}} \right) x \left(\frac{\left[(a_2 - a_3) - (b_2 - b_1\right] x\alpha + (a_3 - b_1)}{6x\sqrt{\Phi_2}} \right) \tag{21}$$

$$(\tilde{C}_{puc})_\alpha = \left(\frac{\left[(a_2 - a_1)x\alpha + a_1\right] - \mu}{3x\sqrt{\Phi_1}} \ , \ \frac{\left[(a_2 - a_3)x\alpha + a3\right] - \mu}{3x\sqrt{\Phi_2}} \right) \tag{22}$$

$$(\tilde{C}_{plc})_\alpha = \left(\frac{\mu - \left[(b_2 - b_3)x\alpha + b_3\right]}{3x\sqrt{\Phi_1}} \ , \ \frac{\mu - \left[(b_2 - b_1)x\alpha + b_1\right]}{3x\sqrt{\Phi_2}} \right) \tag{23}$$

$$\tilde{C}_{pkc} = \min\{\tilde{C}_{puc} \ , \tilde{C}_{plc}\} \tag{23}$$

where; $\Phi_1 = \frac{n\tilde{\sigma}}{(1-\alpha)X^2_{L,0.005} + (\alpha\ x\ n)}$; $\Phi_2 = \frac{n\tilde{\sigma}}{(1-\alpha)X^2_{R,0.005} + (\alpha\ x\ n)}$; X^2_R and X^2_L are the points on the right and left sides of X^2 density (see Buckley, 2004) and $\tilde{C}_{pc}$, $\tilde{C}_{puc}$, $\tilde{C}_{plc}$ and $\tilde{C}_{pkc}$ are fuzzy robust PCIs.

In the study authors also define upper and lower specification limits using trapezoidal fuzzy number. And using a ranking method a comparison of fuzzy PCIs is also performed. After the implementation of the proposed system illustrating obtained results, Kaya and Kahraman concluded that compared to crisp ones, fuzzy analyses for robust PCIs have some advantages as they are more sensitive and include more information than crisp robust PCIs.

The literature exposes that, fuzzy quality control and fuzzy PCA (including fuzzy PCIs) have considerable amount of advantages and remarkable capabilities than their crisp types. They provide more information, they are more sensitive and flexible; and also, more appropriate for implementation to real life cases as they successfully can illustrate human judgment. For these reasons, strongly claiming that "the studies on quality control and PCA in fuzzy domain will rapidly increase" will not be a wrong proposition.

4. Six Sigma methodology

4.1 Relationship between process capability and Six Sigma

The technical elaboration of Six Sigma can be achieved through the use of normal distribution and PCIs. Historically, the creators of Six Sigma employed Cp, as it was accepted as a standard quality measure. Six Sigma was developed for solving the complexity of products and observing the failure of the products in order to achieve the predictive performances (Ramberg, 2002). Similar to Six Sigma methodology, in a process capability study, the number of standard deviations between the process mean and the nearest specification limits is given in sigma units. The sigma quality level of a process can be used to express its capability that means how well it performs with respect to the specification limits. By the way, in terminology of statistics, sigma represents the variation about the process mean. The application of Six Sigma methodology provides reduction in variance and augmentation in the process capability.

As it is mentioned above, a Six Sigma process can be interpreted in terms of process capability, which is associated with process variation by using PCI, such as Cpk. Nowadays, most of the manufacturers are required to produce a product with a specified Cpk value. As the market competition is getting tougher and tougher, organizations are under pressure to sustain world class competition so that they need to meet or exceed this specified Cpk value or quality level. It should be noticed that Cpk values are related to sigma quality levels. Higher value of Cpk indicates a better process. For instance; a process capability, that is, Cpk of 1.00 is roughly equivalent to three sigma capability. That is, the mean plus and the mean minus three standard deviations should be the points at which the nearest specification limits lie. With three sigma capability or Cpk = 1.00, a process will produce approximately 99.73% good product or 0.27% bad product. This represents an unacceptably high level of poor products. On the other hand, nowadays high quality standards dictate reducing variation by four standard deviations between the process mean and the nearest specifications. This corresponds to the value of Cpk = 1.33. At this level, the process will produce approximately 99.9937% good product or 0.0063% bad product. This

represents a better figure than the figure of three sigma capability (or Cpk = 1.00), but it is still having high level of poor products.
Process capability measures have been used to provide number of nonconforming product. As it is mentioned in earlier sections, ppm is used in this regard. At ± 3 sigma level, the probability of producing a product within specification limits is 0.9973. This implies 2700 ppm. Therefore, at a six sigma capability level, a process will produce very few defects. This level represents a Cpk value of 2.0 which is more commonly referred to as six sigma capability.

4.2 Statistical interpretation of Six Sigma

In Six Sigma process, as its name implies, there are six standard deviations between the process mean and specification limits, when the process is centered. The objective of using Six Sigma approach is to reduce process variation, and thereby defects. The six sigma metric uses dpmo, which is the abbreviation for defects per million opportunities. Here, opportunities represent the number of potential chances within a unit for a defect to occur. It is essential to be consistent about the definition of the opportunities because by increasing the number of opportunities over time, a process may be artificially improved (Montgomery, 2009). Computation of dpmo is given below:

$$\text{dpmo} = \frac{\text{total number of defects}}{\text{number of units x number of opportunuties}} \tag{24}$$

Equivalently, dpmo can also be computed like that:

$$\text{dpmo} = \frac{\text{dpu x } 10^6}{\text{opportunuties for error}} \tag{25}$$

where dpu stands for defects per unit. Computation of dpu is given below:

$$\text{dpu} = \frac{\text{total number of defects}}{\text{total number of units}} \tag{26}$$

When dealing with defects or nonconformities, dpu statistic can also be used as a measure of capability. From sample data, quantities of dpu can be estimated, too. Larger samples provide more reliable estimates. Notice that, measure of dpu does not directly take the complexity of the unit into account whereas measure of dpmo does (Montgomery, 2009).
Six Sigma represents a quality level of at most 3.4 dpmo in the long term. Unavoidable assignable causes lead processes to shift 1.5 standard deviations from process mean toward either specification limit that would provide the maximum of 3.4 defects per million. That means Six Sigma measure of process capability allows process mean to shift by up to 1.5 sigma over the long term basis. In order to achieve this goal in the long term, the process capability has to reach the Six Sigma level in the short term, that is, the range between the process mean and the specification limits contains six process standard deviations on both sides of the process mean In this way, the defect rate of a Six Sigma process is only about 0.002 dpmo. However, if the process mean shifts 1.5 process standard deviations over time, the defect rate will increase from 0.002 dpmo to 3.4 dpmo (Feng, 2008). For Six Sigma process, 3.4 dpmo value is the area under the normal curve beyond 6-1.5= 4.5 sigma. Same

logic is valid for three sigma process, that is, 66,807 dpmo value is the area under the normal curve beyond 3-1.5=1.5 sigma (Antony et al., 2005).
As a matter of fact that, Six Sigma has been accepted to mean a 4.5 sigma process, not true Six Sigma process, just because of Six Sigma professionals have allowed for the process to drift by up to 1.5 standard deviations from the process mean. Actually, a process that operates true Six Sigma performance takes up 50% of the specification if the process is centered. This gives Cp = Cpk = 2.00. A process such as this will produce defects at a rate of only approximately 2 parts per billion. From this standpoint, a process with a Cp = 2.00 can have 1.5 sigma drift that is equivalent to 4.5 sigma process. That is, the mean will be 4.5 sigma from the specification limit at the edges of the drift. A 4.5 sigma process yields a 3.4 ppm defect level.
For a process that has a lower quality level than Six Sigma, the success rate will decrease significantly when the process shifts. In this point of view, if an organization is operating at Six Sigma level, it is defined as having less than 3.4 dpmo. This corresponds to a success rate of 99.9997%. On the other hand, if an organization is operating at three sigma level, it is defined as having 66,807 dpmo. This corresponds to a success rate of 93%. (McClusky, 2000). In other words, the fraction outside of the specifications for the three sigma process increases dramatically compared to the fraction for a Six Sigma process and may cause serious quality problems over time. Therefore, three sigma level cannot be regarded as having good quality performance as it is not good enough for many products or processes that attempt to avoid quality problems in the long run.
Literally, Six Sigma is achieved when the process width is half of the specification band. Six Sigma requires process mean is being in control. Inevitably, process mean would not be closer than six standard deviations from the nearest specification limit. That is, Six Sigma needs process specifications are at six standard deviations beyond the process mean. For a Six Sigma process, process potential index Cp and process actual index Cpk would be necessarily 2.00, when process is centered. For a Six Sigma process, actual process performance, that is, Cpk would be 1.5, when there is 1.5 Sigma shift in the process mean.
In general conclusion, Six Sigma is a business approach that drives defects produced by all processes down into parts per million levels of performance as it is accepted as a measure of process performance and the process operating at Six Sigma quality has a defect rate of 3.4 parts per million opportunities (Harry, 1998). In other words, 3.4 dpmo is challenged to be obtained in Six Sigma process. For this reason, Six Sigma is represented by 3.4 defective parts per million. This means it is about improving the process capability for all CTQs from all processes in the organization. The goal in a Six Sigma organization is to achieve defect levels of less than 3.4 ppm for every process in the organization and for every CTQ characteristic produced by those processes.

4.3 Six Sigma methodology

In order to sustain world-class competition, organizations should attain Six Sigma activities by integrating their knowledge of the process with statistics, engineering and project management (Anbari, 2002). Six Sigma is a quality management philosophy as well as a methodology that focuses on reducing variation, eliminating defects and improving the quality of processes, products and services. In other words, Six Sigma Methodology is defined as a data-driven, statistics-based approach and a project-driven management that improves processes, products and services of organization by continuously reducing both

nonconforming items or mistakes and variation as well as costs in the organization. In the literature, Six Sigma has also proven to be a customer-focused and a robust methodology.
In practice, organizations should give importance to improve overall performance instead of detecting and counting defects. The application of Six Sigma methodology provides reduction in variance and augmentation in the process capability, and process performance, simultaneously. Significant improvement in process capability and process performance can be achieved after a successful implementation of Six Sigma methodology that is accepted as a rigorous concept of quality control with this feature.
One of the advantages of the Six Sigma methodology over the other process improvement initiatives is that the use of data analysis tools in Six Sigma projects, which enables to identify process hindering problems and demonstrate the improvements using objective data, accurately. In the literature, several researchers or authors classified the tools and perspectives of Six Sigma methodology in several different ways. For instance; Kwak and Anbari (2006) categorized Six Sigma methodology into two major perspectives, which are statistical and business perspectives. The statistical perspectives of Six Sigma must complement business perspectives and challenge to the organization for a successful implementation of Six Sigma projects. Originally, statisticians created the Six Sigma concept. From the statistical point of view, Six Sigma is defined as having less than 3.4 dpmo. Equivalently, this corresponds to a success rate of 99.9997%. By using statistical tools and techniques, organizations improve sigma quality level as well as process capability, and process performance simultaneously. Feng (2008) highlighted that the requirement of 3.4 dpmo or Cpk of 1.5 is not the ultimate goal of Six Sigma. According to Feng, the attitude is to establish the right business strategy toward organizational excellence. From this standpoint, for the business perspectives of Six Sigma, it is accepted as a business strategy in the business environment that concentrates on improving the effectiveness, efficiency of all operations to meet or exceed customer requirements as well as productivity, business profitability and financial performance (Kwak & Anbari, 2006; Antony & Banuelas, 2001). Beneficial contributions can be expressed in terms of financial returns for organization as Six Sigma increases return on investment (ROI) by process improvement through cost savings as it reduces defects and improves efficiency. Consequently, it results in enhanced customer satisfaction as it fulfills quality requirements. As a result of this, increase in market share can be achieved in the competitive global market.
According to Antony et al. (2005), Six Sigma is a systematic methodology that employs statistical and non-statistical tools and techniques for continuous quality and process improvement and for managing operational excellence. While implementing project-by-project, Six Sigma provides an overall process improvement that clearly shows how to link and sequence individual tools (Feng, 2008). Six Sigma is a strategy for achieving significant financial savings to the bottom-line of the organization. As a matter of fact, organization's ROI can be maximized through the elimination of defects in the processes. Thence, Six Sigma approach is starting with a business strategy and ending with top-down implementation and is having a significant impact on profit by continuously reducing defects throughout the processes of organization and thereby improving customer satisfaction. It must be taken into account that Six Sigma quality level of performance or Six Sigma process capability should not be the primary objective for all the processes. A lower sigma quality level of performance can be acceptable for some processes except the vital ones that are related with zero tolerance for mistakes such as healthcare, safety, reliability, and so on.

According to Allen (2006), tools that are used in Six Sigma methodology can be categorized as tools of statistical methods and quality management, which are very useful in identifying and eliminating causes of defects in business processes by examining the inputs, the outputs, and the relationship between the inputs and outputs.

Tools of statistical methods	Tools of quality management
Statistical Hypothesis Testing	Process Mapping
Regression Analysis	Cause-and-effect diagrams
SPC	Pareto charts
DOE	QFD
ANOVA	FMEA

Table 1. Tools of statistical methods and quality management

In addition to all these tools and techniques, researchers and practitioners observed that Six Sigma has its inherent limitations and cannot be used as a universal solution for any process in any organization. In order to enhance the effectiveness of Six Sigma, additional tools and techniques should be integrated. There is a recent technical development in the fields of management science as well as statistics and engineering which provide more effective tools for enhancing the efficiency and the productivity of organizations such as queuing systems, heuristics, and data envelopment analysis (DEA) (Tang et al., 2007).

Six Sigma builds on improvement methods that have been proved to be effective and integrates the human and process elements of process improvement. The human elements of process improvement consist of teamwork, customer focus and organization's culture change. On the other hand, the process elements of process improvement consist of understanding the types of process variation, process stability, PCA, and DOE for identifying, reducing or eliminating process variation, and thence improve process performance and process capability at the same time (Antony et al., 2005).

Feng (2008) defined Six Sigma as a systematic approach for structured and process-oriented quality or performance improvement and classified two road maps, which are provided by Six Sigma methodology in order to achieve optimum business performance benchmarks for organizations. One is known as the road map for Six Sigma process improvement that is called DMAIC Procedure that consists of five phases, which are Define (D), Measure (M), Analyze (A), Improve (I), and Control (C). DMAIC Procedure involves the improvement of existing processes by removing defects, without changing the fundamental structure of the processes. The other road map is known as Design for Six Sigma (DFSS). DFSS is a Six Sigma approach that involves changing and redesigning the process at the early stages of product or process life cycle. DFSS also consists of five phases, which are Define (D), Measure (M), Analyze (A), Design (D) and Verify (V).

4.4 Key factors for a success of Six Sigma project implementation

In this section, key factors that influence the success of Six Sigma project implementation for improving overall management process would try to be identified. The success of Six Sigma

is related to a set of cross-functional metrics which lead to significant improvements in customer satisfaction and bottom-line benefits (Antony et al., 2005). Generally, wider applications of Six Sigma principles to the organization are achieved through sustained and visible management commitment and involvement as well as whole organizational commitment and organizational infrastructure; organizational cultural refinement; effective project management; continuous education and training, and etc. (Kwak & Anbari, 2006). It should be noticed that, these issues are basically performed with the help of statistics, quality and process improvement tools and techniques.

There can be positive impact on application of Six Sigma when there is continuous managerial support for implementation process. According to Haikonen et al. (2004), managers should adopt as well as internalize Six Sigma philosophy throughout the organization. Top-management involvement and provision of resources and training activities are inevitable for a successful implementation of Six Sigma (Halliday, 2001). Management involvement and organizational commitment are influential to restructure the business and change the attitudes of the organization toward Six Sigma (Hendricks & Kelbaugh, 1998).Commitment of resources, time, money and effort from entire the organization is essential for Six Sigma project implementation. Organizational infrastructure needs to be established with well trained individuals. Before introducing Six Sigma concepts and tools, SWOT analysis can be performed in order to identify strengths and weaknesses of organization to ensure long term sustainability of Six Sigma Methodology (Kwak & Anbari, 2006).

Refining the organizational culture continuously is also compulsory for a successful implementation of Six Sigma. Leadership are necessary for a change in organizational culture.The attitudes of the employees or all of the participants should also be changed towards the Six Sigma philosophy. More concisely, implementation of a Six Sigma program needs the right mindset and attitude in the people working at all levels within the organization (Antony & Banuelas, 2001). For this purpose, clear communication plan needs to be developed. Motivation and education for Six Sigma are influential factors for refining the organizational culture, too. It should be taken into consideration that organizational cultural changes require time and commitment. Effective Six Sigma principles as well as practices can be more likely achieved by refining the organizational culture continuously (Kwak & Anbari, 2006).

Six Sigma project selection, review and tracking are fundamental parts of effective project management. Effective project management includes careful consideration of projects to be feasible, organizationally and financially beneficial and conformation of appropriate set of measures and metrics to satisfy customer requirements. Periodic review of project evaluates the state of the project and performance of Six Sigma tools and techniques. Documentation is necessary for tracking of projects within project constraints that are mainly cost, time and quality (Kwak & Anbari, 2006).

Continuous education and training give a clear sense for participants for understanding the tools and techniques and principles of Six Sigma Methodology. On this account, the implementation of Six Sigma should start with the training of a dedicated workforce and the education across the organization. It should be considered that there can be inherent drawback of misapplication of Six Sigma Methodology when personnel are trained inadequately. Although Six Sigma is deployed from top down, people in the organization

need necessary training to realize Six Sigma improvement and its potential benefits to the organization and themselves. In order to implement Six Sigma tools and techniques effectively, communication techniques should be widespread throughout the organization. Participants should be well informed about the Six Sigma tools and techniques and communicate with actual data analysis. Identifying key roles and responsibilities of participants for implementing Six Sigma project should be well defined. Learning the principles behind the Six Sigma methodology requires Six Sigma training activities. Training should also cover quantitative and qualitative measures and metrics along with leadership and project management. Training is a key success factor in implementing Six Sigma projects (Kwak & Anbari, 2006). Understanding of Six Sigma methodology accompanied by tools and techniques is very important for successful Six Sigma applications.

Historically, statisticians created Six Sigma concept, thus, the origin of Six Sigma comes from statistics. In this connection, the success of Six Sigma project implementation for improving overall management process is absolutely related to the appropriate usage of tools and techniques of statistics and quality. By utilizing statistical tools and techniques, Six Sigma methodology enables practitioners to identify process hindering problems accurately. Also, utilization of statistical tools and techniques demonstrate the improvements based on usage of objective data. That's why Six Sigma is accepted as a data-driven approach as it needs to quantify the process by using actual data. Therefore, statistical thinking is vital for Six Sigma methodology, reduction of defects and variation. As Six Sigma originated from the statistical concept for quality improvement, the role of management in statistical thinking is important for quality and process improvement efforts.

5. Lean Six Sigma

Lean Six Sigma is a combination of concepts of two productivity improvement programs, which are Six Sigma and Lean Manufacturing. Particularly, Six Sigma is a quality management philosophy as well as a methodology that focuses on reducing variation, defects and improving the quality of processes, products, and services. Six Sigma cannot reduce waste or reduce cycle time in processes alone. On the other hand, Lean Manufacturing is a methodology that focuses on reducing waste and cycle time in processes. Lean cannot reduce variation alone. To sum up, Lean Six Sigma is an approach that focuses on improving quality by reducing variation and defects as in Six Sigma and eliminating waste along with reducing cycle time in an organization as in Lean Manufacturing.

According to George (2002) Lean Six Sigma is a methodology that maximizes shareholder value by achieving the fastest rate of improvement in customer satisfaction, cost, quality, process speed and invested capital. In order to eliminate waste and reduce variation in any process, Lean Six Sigma can be used.

In fact, Six Sigma differs from Lean Manufacturing because they attack different types of problems. Basically, Six Sigma is concerned with less visible problems in processes such as variation in performance. Six Sigma tools require advanced training and expertise of specialists. However, Lean Manufacturing is concerned with visible problems in processes such as inventory, material flow and safety. Lean tools are more intuitive and easier to apply. Organizations are recommended to start with basic lean principles and evolve toward more sophisticated Six Sigma tools and techniques.

6. Conclusions and recommendations

In today's competitive business environment, the competition power of small and medium size enterprises, companies and even countries (either in private, public or military sectors) in the national and international business area are mainly based on customer (internal or external) relations; understanding the needs of customer, ability and flexibility of immediate response to needs of customer and requirements for providing capability to fulfill those needs. All activities performed to provide capabilities for satisfying customer needs include many sophisticated interrelated functions and processes either directly or indirectly based on what customer wants or more specifically, the customer demand; such as decision making, management, new product development, production, marketing, logistics, finance, quality control, human resources and etc.; which all together compose dynamic, complex and chaotic structures. These of complex structures with all interrelated functions have to be designed and managed perfectly pointing us to two well-known terms supply chain networks (ScNs) and ScN management. In such dynamic and complex systems (i.e., ScN like systems), all processes have to be managed successfully; which in fact, can only be achieved with on time, true and appropriate control mechanisms. As presented through the whole chapter, basic tool for establishing such mechanisms are PCA. PCA and fuzzy PCA together with their integral parts PCIs and fuzzy PCIs occupy a vital place in every field where computational controls are needed. In industrial practice PCA and fuzzy PCA; which is mainly used for predicting how well the process will hold the tolerances, can be used in many segments of the product cycle. They can be used in production and production planning as it reduces the variability in a process and plans the sequence of production processes when there is an interactive effect of processes on tolerances. Also; in process and product design, by assisting designers in selecting or modifying a process and in specifying the performance requirements PCA and fuzzy PCA can successfully be used. Even in the selection of competing suppliers, PCA and fuzzy PCA play important role.

Six Sigma also positively impacts many CTQ features such as timeliness/speed, cost, and quality of product or service as it identifies root causes and eliminates variations and defects. After a successful implementation of Six Sigma project, savings from reduced rework, less waste and decrease in customer returns can be obtained. So, this approach is also one of the indispensible in today business environment. Due to its importance, with every passing time, this approach is developed with researches and ideas, like lean approaches as mentioned before.

As a result, process capability and six sigma methodology; including fuzzy and lean approaches plays an important role in daily and theoretical scientific life. Today it would not be wrong to claim that it is a must for every enterprise in every field to adopt these modalities into their activities without wasting time. It may also be concluded that, studies on these significant subjects on fuzzy domain is still relatively unrefined. In the future, much more effort will and must be expend on these subjects in fuzzy domain.

7. References

Allen, T. T. (2006). *Introduction to Engineering Statistics and Six Sigma: Statistical Quality Control and Design of Experiments and Systems*, Springer, ISBN: 1852339551, London.

Anbari, F. T. (2002). Six sigma method and its applications in project management. *Proceedings of the Project Management Institute Annual Seminars and Symposium*, San Antonio, Texas, Oct 3-10, Project Management Institute, Nextown Square, PA.

Antony, J. & Banuelas, R. (2001). A strategy for survival. *Manufacturing Engineer*, Vol.80, No3, 119-121, ISSN: 0956-9944.

Antony, J., Kumar, M., & Tiwari, M. K. (2005). An application of six sigma methodology to reduce the engine-overheating problem in an automotive company. *Journal of Engineering Manufacture*, B8, 14, 633-646, ISSN: 2041-2975.

Automotive Industry Action Group (AIAG) (2002). *Measurement system analysis*, (3rd ed.) Southfield, MI: Author.

Bellman, R. E., Zadeh L.A., (1970). Decision- making in fuzzy environment, *Management Science*, Vol.17, 141-164.

Bothe, D. R. (2002). Discussion, *Journal of Quality Technology*, Vol. 34, No.1, ISSN: 0022-4065.

Buckley, J. J. & Eslami, E. (2004). Uncertain probability II: The continuous case, *Soft Computing*, Vol. 8, 193-199.

Buckley, J. J. (2005a). Fuzzy statistics: Hypothesis testing, *Soft Computing*, Vol. 9, 512-518.

Buckley, J. J. (2005b). Fuzzy statistics: Regression and prediction, *Soft Computing*, Vol. 9, 769-775.

Deleryd, M. (1999). The effect of skewness on estimates of some process capability indices. *International Journal of Applied Quality Management*, Vol. 2, No. 2, 153-186, ISSN: 1096-4738.

Chen, K. S., Chen, T. W., (2008). Multi-process capability plot and fuzzy inference evaluation, *Int. Journal of Production Economics*, Vol. 111, 70-79.

Chen, T. W., Chen K. S., Lin J. Y. (2003). Fuzzy evaluation of process capability for bigger-the-best type products, *Int. Journal of Advanced Manufacturing Technology*, Vol. 21, 820-826.

English, J. R. & Taylor G. D. (1993). Process capability analysis- a robustness study. *International Journal of Production Research*, Vol. 31, No. 7, 1621-1635, ISSN: 0020-7543.

Evans, J. R. & Lindsay W. M.(2008). *The Management and Control of Quality*, (7th ed.), Thomson, ISBN: 0324382273, OH USA.

Faraz A. & Moghadam M. B. (2007), Fuzzy control charts a better alternative for shewhart average charts, *Quality and Quality*, Springer, NJ USA.

Feng, Q.(2008). Six Sigma : Continuous Improvement Toward Excellence (Chapter 3) In: *Collaborative Engineering: Theory and Practice,* A.K. Kamrani, E.S. Abouel Nasr (eds.), 43-60, Springer, ISBN: 978-0-387-47319-2, NY, USA

Franklin LeRoy A. (1999). Sample size determination for lower confidence limits for estimating process capability indices. *Computers and Industrial Engineering*, Vol.36, 603-614.

George, M. (2002). *Lean Six Sigma Combining Six Sigma Quality with Lean Speed*, McGraw-Hill, ISBN: 0071385215, USA.

Gulbayi, M. & Kahraman C. (2006). Development of fuzzy process control charts and fuzzy unnatural pattern analyses, *Computational Statistics and Data Analysis*, Vol. 51, 434–451.

Gulbay, M. & Kahraman C. (2007). An alternative approach to fuzzy control charts: direct fuzzy approach, *Information Sciences*, Vol. 177, 1463–1480.

Haikonen, A., Savolainen, T., & Jarvinen, P. (2004). Exploring six sigma and CI capability development: preliminary case study findings on management role. *Journal of Manufacturing Technology Management*, Vol.15, No.4, 369-378.

Halliday, S. (2001). So what exactly is Six Sigma*? Works Management*, Vol. 15, No.1, 15.

Harry, M. J. & Lawson, J. R. (1992). *Six sigma producibility analysis and process characterization*, Addison-Wesley Pub., ISBN: 0201634120, USA.

Harry, M. J. (1998). Six Sigma: a breakthrough strategy for profitability, *Quality Progress*, Vol.31, No. 5, 60-64.

Hendricks, C.A. & Kelbaugh, R. (1998). Implementing Six Sigma at GE, *The Journal of Quality and Participation*, Vol.21, No.4, 48-53.

http://wordnetweb.princeton.edu/, (2009).

Kahraman, C. (2006). *Fuzzy Applications in Industrial Engineering*, Studies in Fuzziness and Soft Computing, Vol. 201, Springer Verlag , NJ USA.

Kahraman, C. & Kaya, I. (2008). Fuzzy process capability indices for quality control of irrigating water, *Stochastic Event Research and Risk Assessment*.

Kahraman, C. & Kaya, I. (2009). Fuzzy process accuracy index to evaluate risk assessment of drought effects in Turkey, *Human and Ecological Risk Assessment: An International Journal, Vol. 15, No. 4,* 789-910.

Kaya, I. & Kahraman, C. (2007). Air pollution control using six-sigma approach, *Proceedings of International Conference on Risk Analysis and Crisis Response,* 100-115.

Kaya, I. & Kahraman, C. (2008a). Fuzzy process capability analyses: An application to teaching process, *Journal of Intelligent and Fuzzy Systems,* Vol. 19, *259*--272.

Kaya, I. & Kahraman, C. (2008b). Fuzzy robust process capability indices for risk assessment of air pollution, *Stochastic Environmental Research and Risk Assessment*.

Kaya, I. & Kahraman, C. (2009). Air pollution control using fuzzy process capability indices in six-sigma approach, *Human and Ecological Risk Assessments: An international Journal,* Vol. 15, No.4, 689-713.

Kaya, I. & Kahraman, C. (2010). A new perspective on fuzzy process capability indices: Robustness, *Expert Systems with Applications,* Vol. 37, 4593-4600.

Klir, G. J., Yuan B.(1995). *Fuzzy sets and fuzzy logic: Theory and applications*, Prentice Hall, ISBN: 0131011715.

Kotz S. & Johnson N.L. (2002). Process capability indices-A review, 1992-2000 (with subsequent discussions and response*), Journal of Quality Technology*, Vol.34, No.1, 2-53.

Kwak, Y. H. & Anbari, F. T. (2006). Benefits, obstacles, and future of six sigma approach, *Technovation*, Vol.26, No. 5, 708-715.

Lee, Y. H., Wei, C.C. & Chang, C.L., (1999). Fuzzy design of process tolerances to maximize process capability, *International Journal of Advanced Manufacturing Technology*, Vol.15, 655-659.

Lee, H. T. (2001). Cpk index estimation using fuzzy numbers, *European Journal of Operational Research*, Vol. 129, 683-688.

McClusky, R. (2000). The Rise, fall, and revival of six sigma, *Measuring Business Excellence*, Vol.4, No 2, 6-17.

Monden, Y. (1993). *Toyota Production System, An Integrated Approach to Just-In-Time*, (2th ed.), Industrial Engineering and Management Institute Press, Norcross GA.

Montgomery, D. C (2009*). Statistical Quality Control- A Modern Introduction*, Wiley., ISBN: 978047233979, USA.

Qianmei Feng (2008). Six sigma: Continuous improvement toward excellence. In: *Collaborative Engineering*, A.K. Kamrani, E.S. Abouel Nasr (eds.), 43-60, Springer, ISBN: 9780387473192.

Parchami, A., Mashinchi, M., Yavari, A. R. & Maleki, H. R. (2005). Process capability indices as fuzzy numbers. *Austrian Journal of. Statistics* , Vol. 34 , 391-402.

Parchami, A. & Mashinchi, M. (September 2006). Making decision to evaluate fuzzy process capability index. *Asian Fuzzy Systems Society International Conference (AFSS 2006)* Boading, China, 28-33.

Parchami, A. & Mashinchi, M. (2007). Fuzzy estimation for process capability indices. *Information Sciences* , Vol.177 ,1452-1462.

Parchami, A. & Mashinchi, M. (2008). Testing the capability of fuzzy processes. *Quality Technol. Quant. Management.* — Accepted

Parchami, A. , Mashinchi, M. & Maleki, H. R. (2006). Fuzzy confidence interval for fuzzy process capability index, *J. Int. Fuzzy Syst.*,Vol.17 , 287-295.

Parchami A., Mashinchi M. & Sharayei A. (2010a), An effective approach for measuring the capability of manufacturing processes, *Production Planning and Control*, Vol. 21 250-257.

Parchami A. & Mashinchi M. (2010a), An effective approach for measuring the capability of manufacturing processes, *Production Planning and Control*, Vol. 37, 77-89.

Palit, A. K., Popovic D., (2005). *Computational intelligence in time series forecasting: Theory and engineering applications*, Advances in Industrial Control, Springer Verlag, NJ USA.

Ramberg, J. S. (2002). Discussion. *Journal of Quality Technology*, Vol.34, No.1.

Ramirez B. & Runger G. (2002) Quantitative Techniques to evaluate process stability, Quality Engineering, Vol.18, 53-68.

Ross, T. J. (2004).*Fuzzy Logic with Engineering Applications*, (2th Ed.), Wiley, ISBN: 9780470860748, USA.

Spiring, F., Cheng, S., Yeung A., Leung, B. (2002). Discussion, *Journal of Quality Technology*, Vol.34, No.1.

Tanaka, K. (1997). *An introduction to fuzzy logic for practical application*, Springer, ISBN: 9780387948072.

Tang, L.C., Goh T.N., Lam S.W., Zhang C.W. (2007). Fortification of six sigma: Expanding the DMAIC toolset, *Quality and Reliability Engineering International*, Vol.23, No.1, 3-18.

Tozan, H. Vayvay, Ö. (2008), Fuzzy forecasting applications on supply chains, *WSEAS Transactions on Systems*, Vol.7, 600-609.

Tozan, H. Vayvay, Ö. (2009), A hybrid Grey & ANFIS approach to bullwhip effect in supply chain networks, *WSEAS Transactions on Systems*, Vol.7, 600-609.

Wang, C. H. & Raz, T. (1990). On the Construction of Control Charts Using Linguistic Variable, *International Journal of Production Research*, Vol. 28, No. 3, 477-487.

Yonting, C. (1996). Fuzzy quality and analysis on fuzzy probability. *Fuzzy Sets and Systems*, Vol. 83, 283-290.

Zadeh,L. A. (1965). Fuzzy sets, *Information and Control*, Vol. 8, 338-353.

10

Adaptive Involvement of Customers as Co-Creators in Mass Customization

Igor Fürstner and Zoran Anišić
Subotica Tech – College of Applied Sciences
Serbia

1. Introduction

In the twenty-first century, a company has to organize around the customer in order to be a successful and viable firm. Today, the marketplace is customer driven. Customers expect to get what they would like, with a side order of customization. This approach raises several questions that have to be answered, one of which is that despite nowadays customers are knowledgeable in general, they are still far from being experts that can really co-create a product or a service (Galbraith, 2005). Companies are forced to change their activities from a seller's point of view towards a buyer's point of view, which results in a drastic increase of product variety offered by enterprises. That is one of the main characteristic trends of the modern economic system (Forza & Salvador, 2007). To maintain their competitiveness, companies are modularizing their products and introducing platform concepts, and this transfer from no customizable products to modular products that involve individual customer variants is one of the most important industrial strategies nowadays. The recent development of IT technology enables the software based product configuration systems that support the process of customized product development. They compose customer-specific solutions using the modules based on the customer's requirements.
These drastic changes in modern economy introduce mass customization that alters traditional product development and moves towards a two-stage model, the first, the realm of company/designer establishing the solution space and the second, that of the customer as co-designer. This second stage fundamentally changes the role of the customer from the consumer of a product, to a partner in a process of adding value (Reichwald et al., 2004).
This alteration of traditional product development through the involvement of the customer into the configuration of the final product faces some obvious problems. The fundamental challenge is to avoid the abortion of the configuration process by the customer. In many cases, the customer aborts the configuration process by himself. Major problem areas include the lack of a customer-desired option value regarding a specific attribute within the system as well as the inability of the customer to create definite preferences between certain option values. As a result, the customer aborts the configuration process and does not come up to the sales phase (Hansen et al., 2003). Also if customers are overwhelmed by the configuration task, there is a chance that they may abort the configuration process. Customers usually only want the product alternatives that exactly meet their requirements. If too much of a choice is offered, customers can feel frustrated or confused, and therefore incapable of making proper decisions. This overload of information is sometimes called

external complexity. This external complexity is caused by the limited information processing capacity of humans, the lack of customer knowledge about the product, and customer ignorance about his or her real individual needs (Blecker & Abdelkafi, 2006). Based on problem analysis regarding customers' involvement in the configuration process, the main areas of investigation to be considered are the minimization of the complexity experienced by the customers (Berger & Piller, 2003; Kumiawan et al., 2003) and the reduction of the cognitive overhead, considering not only the extent of choice, but also the lack of understanding about which solution meets their needs and also the uncertainties about the behavior of the supplier and the purchasing process (Franke & Piller, 2003).
The chapter presents one approach in solving the presented problem, by the introduction of a methodology for adaptive involvement of customers as co-creators in mass customization of products and services.

2. Customer profile configuration

Generally, the identification and implementation of customer requirements are significant issues for successful product development (Engelbrektsson & Soderman, 2004). To be able to select or filter objects for an individual, information is needed about the individual (Levy & Weld, 2000; Schubert & Koch, 2002). Based on experience, the problem of adapting the process of co-creation to different customers can be solved by identification of different customer profiles that suit each individual customer's needs and limitations.
The area of customer profiles (Fig. 1) consists of general information about customers, which usually deals with basic and demographic attributes, information about specific product interests, information about general interests, information about relationships to other customers, information about the buying history and usage/interaction behavior and ratings of products, product components and certain attributes (Leckner & Lacher, 2003), specific information about customers, which is derived from input questions (Čović et. al, 2009; Fürstner & Anišić, 2009a; Maravić et al., 2009) and contextual information about customers, such as time of the day, the date, etc. (Schubert & Koch, 2002; Koch & Moeslein, 2003).

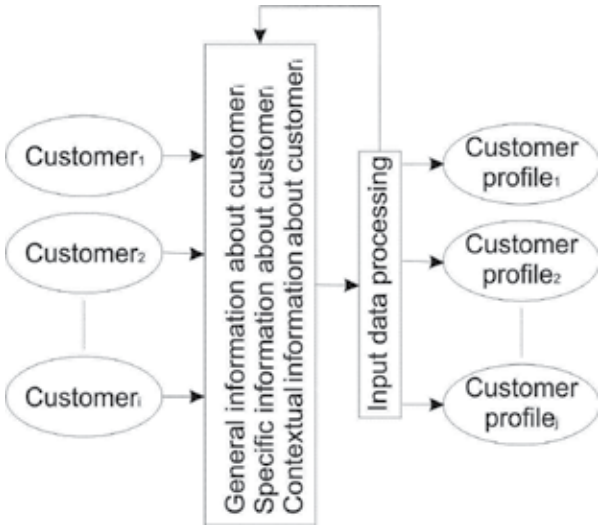


Fig. 1. Area of customer profiles

Voices of customers should not only be elicited at the front-end of the process, but rather frequently at various junctures along the process (Chong et al., 2009). Besides, not all the requirements can be known at the outset of the task (Gero & Kannengiesser, 2004). It is therefore necessary to collect customer opinions consistently.

The rest of the chapter deals with collecting and using specific information about customers, while general and contextual information about customers are not discussed here.

To configure the appropriate customer profile, specific information about customers is needed. Therefore a set of initial questions is asked at the beginning of the co-creation process.

There is a need to analyze the answers generated by each customer and to use them to form a customer profile. A number of approaches from the field of data analysis may be used, nevertheless the nature of the questions and the answers refer to the use of a non-crisp logic; therefore fuzzy logic is used to determine the appropriate customer profile (Zimmermann, 1988; Hanss, 2005; Bojadziev & Bojadziev, 2007).

The initial development of the theory of fuzzy sets was motivated by the perception that traditional techniques of systems analysis are not effective in dealing with problems in which the dependencies between variables are too complex or too ill-defined to admit of characterization by differential or difference equations. Such problems are the norm in biology, economics, psychology, linguistics, and many other fields. A common thread that runs through problems of this type is the unsharpness of class boundaries and the concomitant imprecision, uncertainty and partiality of truth. The concept of a fuzzy set is a reflection of this reality (Bojadziev & Bojadziev, 2007).

Generally, a fuzzy number A (Fig. 2) is defined on the universe R as a convex and normalized fuzzy set, by a membership function $\mu_A(x)$ (Hanss, 2005; Bojadziev & Bojadziev, 2007).

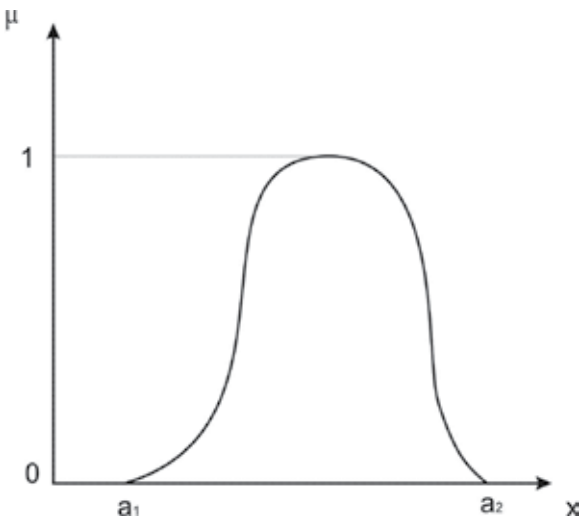


Fig. 2. Fuzzy number A

The adaptive customer profile configurator is discussed by using trapezoidal (triangular) fuzzy numbers (Fig. 3). A trapezoidal fuzzy number A is defined on R by a following membership function (1).

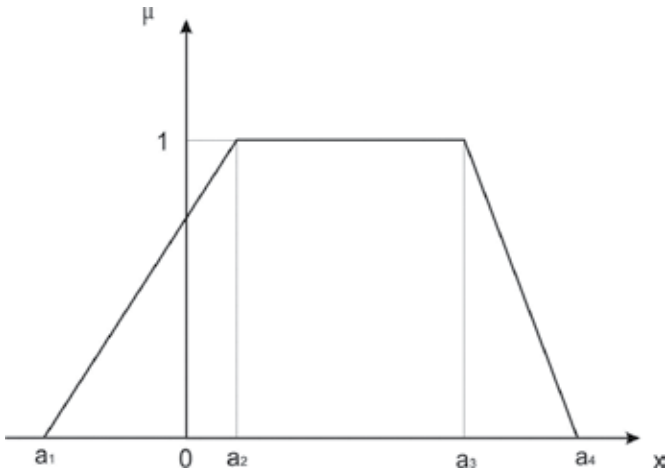


Fig. 3. Trapezoidal fuzzy number

$$\mu_A(x)=\begin{cases}0, x<a_1\\ \dfrac{x-a_1}{a_2-a_1}, a_1\le x\le a_2\\ 1, a_2\le x\le a_3\\ \dfrac{x-a_4}{a_3-a_4}, a_3\le x\le a_4\\ 0, x>a_4\end{cases} \quad (1)$$

Each question from the set of initial questions can have answers that can range from 0 to 1. 0 usually means that the answer is negative, 1 means that the answer is positive. Not only the answers are evaluated, but also the order of answering to questions. Also, during and after the process of co-creation, the customer's feedback considering his satisfaction with a configured profile is analyzed and the profile is adapted according to the feedback.

Based on asked questions and answers, several linguistic variables are defined, that can have different values. Next example shows a linguistic variable a that have three values (high, average and low) with the appropriate membership functions $\mu(x)$ for the variable (Fig. 4).

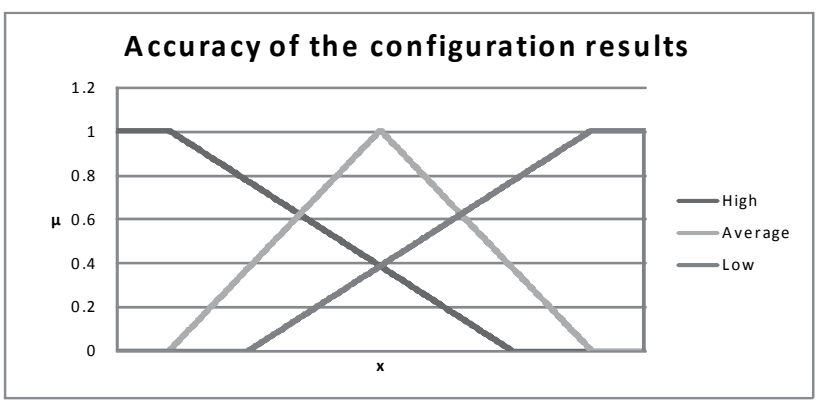


Fig. 4. Values for the linguistic variable

It was mentioned before that during the process of customer profile configuration, the order of answering the questions is also taken into consideration. The reason for doing so is that customers usually, based on their belief, sooner answer questions that are of higher importance to them than questions that are not. There is also a possibility that customers do not answer unimportant questions at all; then the value of the answer is 0.5 (Chen, 2009).
For the same answer values (customer input), the membership functions change, based on the answering order. For the same variable, if the answer to the question is the first one, the membership functions taper, i.e. the equations are changed in the following manner (2).

$$\begin{aligned} \mu_{a=high}{}^{1st}(x) &= \left[\mu_{a=high}(x)\right]^{y_{high}} \\ \mu_{a=average}{}^{1st}(x) &= \left[\mu_{a=average}(x)\right]^{y_{average}}, \quad y_i \geq 1 \\ \mu_{a=poor}{}^{1st}(x) &= \left[\mu_{a=poor}(x)\right]^{y_{poor}} \end{aligned} \tag{2}$$

It results in a more unique response (Fig. 5).

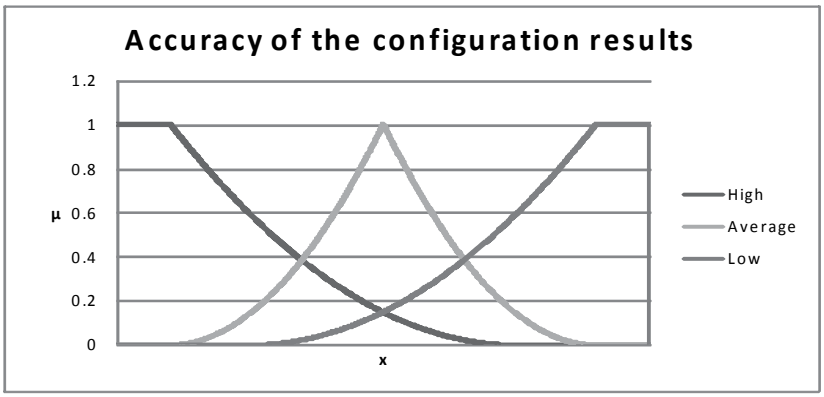


Fig. 5. Values for the linguistic variable

If the answer to the question is the last one, the membership functions expand, i.e. the equations are changed in the following manner (3).

$$\begin{aligned} \mu_{a=high}{}^{last}(x) &= \left[\mu_{a=high}(x)\right]^{y_{high}} \\ \mu_{a=average}{}^{last}(x) &= \left[\mu_{a=average}(x)\right]^{y_{average}}, \quad y_i \leq 1, \\ \mu_{a=poor}{}^{last}(x) &= \left[\mu_{a=poor}(x)\right]^{y_{poor}} \end{aligned} \tag{3}$$

It results in a more vague response (Fig. 6).
The fuzzy output from the system, i.e. the decision is made in a manner that for i initial questions, each of which can have y_i values, $y_1 * y_2 * ... * y_i$ if-then rules can be defined. The rules are designed to produce j different outputs o_j with defined membership functions (Fürstner & Anišić, 2009b; Fürstner & Anišić, 2009c). Next example (Fig. 7) shows an output with three possibilities defined by the following membership functions (4).

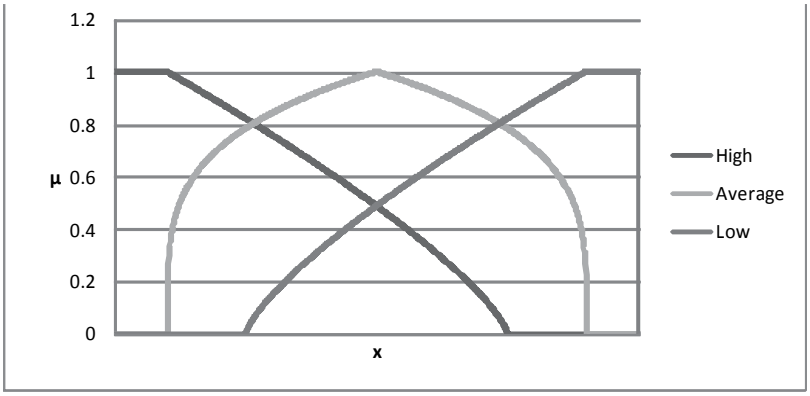


Fig. 6. Values for the linguistic variable

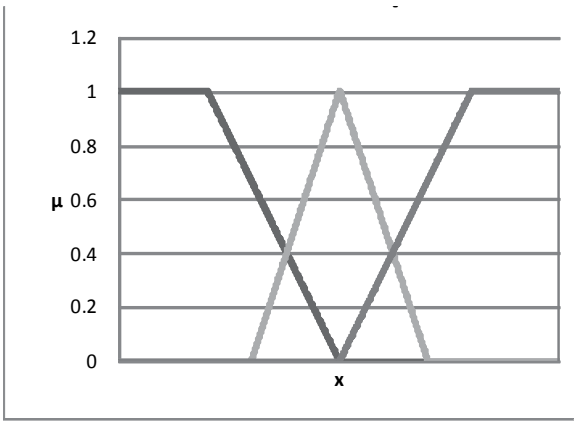


Fig. 7. Output

$$\mu_{o1}(x)=\begin{Bmatrix} 1, & 0\le x\le \alpha_0 \\ \dfrac{\beta_0-x}{\beta_0-\alpha_0}, & \alpha_0<x\le\beta_0 \\ 0, & \beta_0<x\le 1 \end{Bmatrix}$$

$$\mu_{o2}(x)=\begin{Bmatrix} 0, & 0\le x\le \chi_0 \\ \dfrac{x-\chi_0}{\delta_0-\chi_0}, & \chi_0<x\le\delta_0 \\ \dfrac{\varepsilon_0-x}{\varepsilon_0-\delta_0}, & \delta_0<x\le\varepsilon_0 \\ 0 & \varepsilon_0<x\le 1 \end{Bmatrix}, \tag{4}$$

$$\mu_{o3}(x)=\begin{Bmatrix} 0, & 0\le x\le \phi_0 \\ \dfrac{x-\phi_0}{\phi_0-\varphi_0}, & \phi_0<x\le\varphi_0 \\ 1, & \varphi_0<x\le 1 \end{Bmatrix}$$

where $\alpha_0, \beta_0, \chi_0, \delta_0, \varepsilon_0, \phi_0, \varphi_0$ are the initial values.

After the evaluation of if-then rules, an aggregated output is generated. Changes in input membership functions influence the customer profile configuration. For the same answers, but for a different answering order, the configured customer profile can be different (Fig. 8).

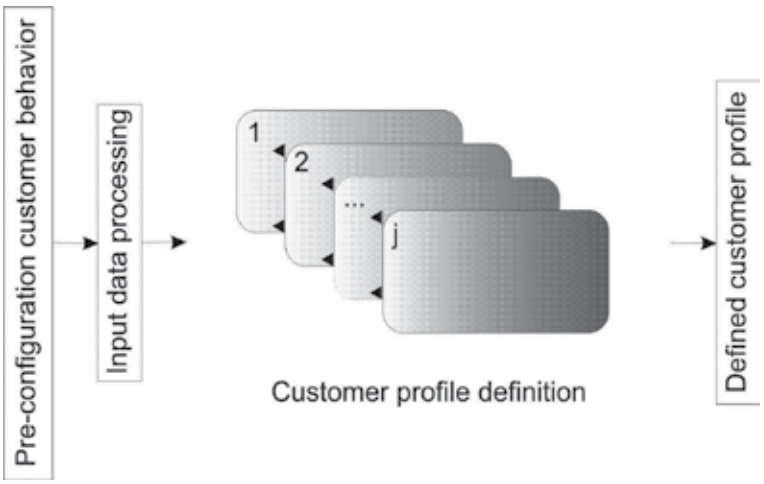


Fig. 8. Customer profile definition

After the configuration task is finished, a feedback is generated. The customer is asked to answer a new set of questions. Each question from the set can have answers that can range from -0.5 to 0.5. -0.5 usually means that the answer is negative, 0.5 means that the answer is positive. The answers to the questions are the feedback about how well the configurator has been adapted to customer's needs and limitations. Initially, all the answers are set to the value of 0, which means that the customer is satisfied with the configuration process.

Based on the answers to questions, the values for input linguistic variables (for example for linguistic variable a) are modified to new values (for example to linguistic variable a_{new}) in the following manner (5).

$$a_{new} = a + \frac{feedback}{2}, \quad 0 \le a_{new} \le 1 \tag{5}$$

This is the input for a new fuzzy output from the system, i.e. a new decision. This new output (o_{new}) takes into consideration whether a customer is satisfied with a configured customer profile. Based on the difference between an original and a new output, the membership functions for o_{i+1}, where o_{i+1} is the output in the future, are shifted left or right to better articulate the future customers' preferences (Fig. 9).

The amount of shifting (sa) is calculated in the following manner (6).

$$sa = \frac{o - o_{new}}{10} \tag{6}$$

The division by 10 is used to assure that the shift is not too big. The resulted customer profile generation algorithm is shown in Fig. 10.

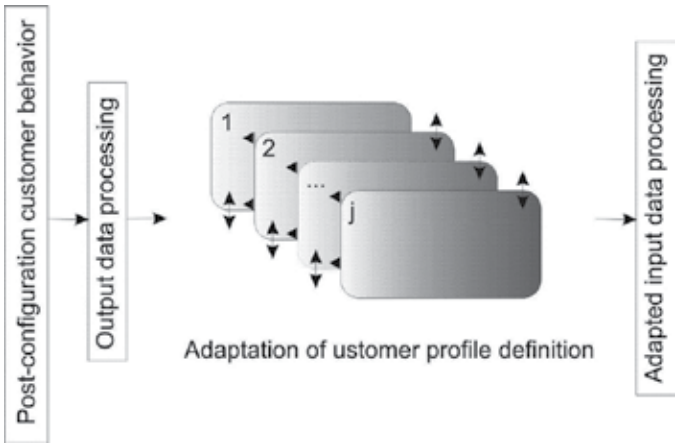


Fig. 9. Adapted input data processing

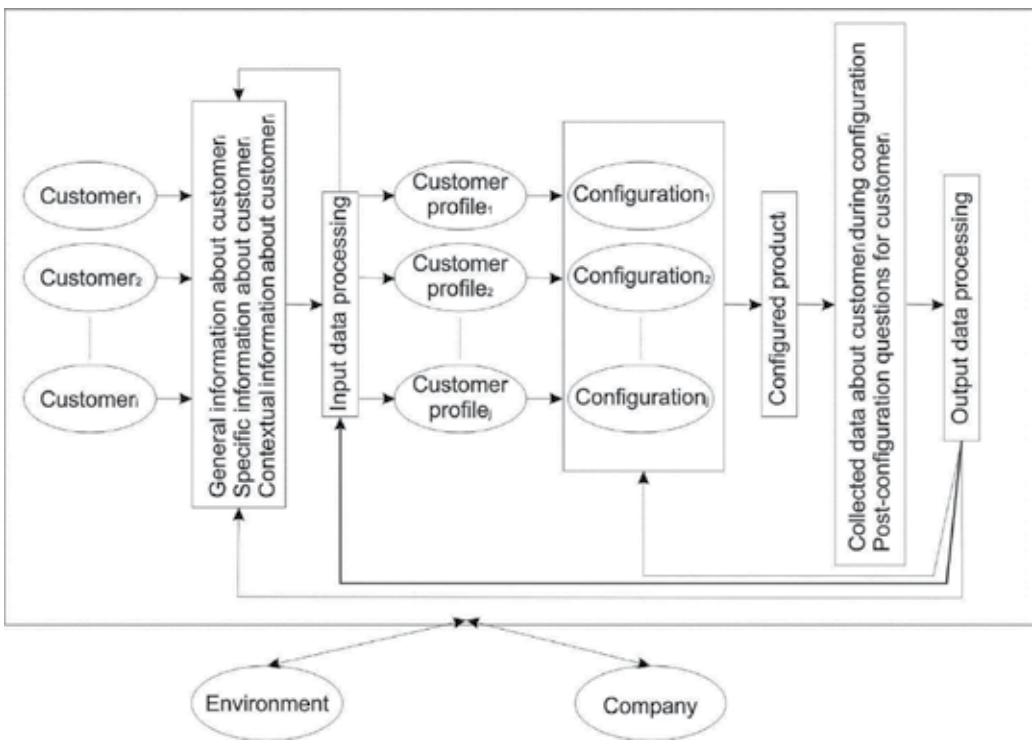


Fig. 10. Customer profile generation

3. Case study

The customer profile generation and the usage of different customer profiles in the process of co-creation are tested on a developed configurator for thermal insulation of buildings with the following characteristics (Fürstner & Anišić, 2009a; Fürstner & Anišić, 2009b, Fürstner & Anišić, 2009c):

- The configurator has to offer web based on-line instant results;
- The result should be based on the latest results in research and practice;
- The configurator should configure customized results, based on the specific characteristics of individual buildings;
- The configurator has to minimize the potential complexity experienced by the customer, by reduction of cognitive overhead;
- The configurator has to be used by professionals, retailers and end customers without specific technical knowledge about thermal insulation;
- The configurator should offer an accurate enough result, which is acceptable in the research field;
- The configurator has to raise the awareness about the necessity and the advantages of proper thermal insulation.

The algorithm for determination of thermal insulation of buildings is not discussed in this paper.

Previously developed configurator that was meant to be used both by customers with average or no technical knowledge and by professionals with proper technical knowledge in the related field of investigation had some limitations, because some of the previous non-professional customers had found the product configurator too complex to use. On the other hand some of the professional customers have found that the configurator lacked the possibility of defining exact and precise input parameters. Other problems included the need for more or less accurate results, as well as more or less time-consuming configuration. These problems were solved by identification of three different customer profiles:

- "Dummy" customer;
- Intermediate customer;
- Professional customer.

The "Dummy" customer is a customer without proper technical knowledge about thermal insulation, or maybe a customer with no need for highly accurate results, or a customer with a need of a fast enough result, etc. The Intermediate customer is a customer with average technical knowledge about thermal insulation, but can also be a customer without proper technical knowledge about thermal insulation but with more time for completing the configuration process or with a need for more accurate result, etc. The Professional customer is a customer with proper knowledge about the problem of thermal insulation; it may also be a customer with average technical knowledge about thermal insulation but with more time for completing the configuration process or with a need for more accurate result, etc.

To configure the appropriate customer profile, three initial questions are asked before the start of the configuration process:

- What is your estimate about your knowledge about thermal insulation?
- What are your needs considering the accuracy of the configuration results?
- How much time do you have for completing the configuration process?

The answers can range from "I have no knowledge about thermal insulation at all" (Where the value of the answer is 0) to "I am a professional in the field of thermal insulation" (Where

the value of the answer is 1) for the first question; from "I need as accurate result as possible" (Where the value of the answer is 0) to "I just want a rough estimate" (Where the value of the answer is 1) for the second question; and from "I have enough time for completing the configuration process" (Where the value of the answer is 0) to "I have limited time for completing the configuration process" (Where the value of the answer is 1) for the third question. Initially, all the answers are set to the value of 0.5. The answers are used as input data for customer profile configuration.

Based on asked questions and answers, three linguistic variables are defined:

- Knowledge about thermal insulation (k), whose values are: very poor, poor, average, good and very good;
- Accuracy of the configuration results (a), whose values are: high, average, low;
- Time for the configuration process (t), whose values are: enough, average, not enough.

The membership functions for the variables are triangular or trapezoidal, and are chosen based on previous testing and experience (7), (8), (9), where the variables are described on the operating domain of $x = [0,1]$ (Fürstner & Anišić, 2009b; Fürstner & Anišić, 2009c).

$$\mu_{k=very\_poor}(x) = \begin{cases} 1, & 0 \le x \le 0.05 \\ \dfrac{0.5-x}{0.5-0.05}, & 0.05 < x \le 0.5 \\ 0, & 0.5 < x \le 1 \end{cases}$$

$$\mu_{k=poor}(x) = \begin{cases} \dfrac{x}{0.3}, & 0 \le x \le 0.3 \\ \dfrac{0.6-x}{0.6-0.3}, & 0.3 < x \le 0.6 \\ 0, & 0.6 < x \le 1 \end{cases}$$

$$\mu_{k=average}(x) = \begin{cases} 0, & 0 \le x \le 0.2 \\ \dfrac{x-0.2}{0.5-0.2}, & 0.2 < x \le 0.5 \\ \dfrac{0.8-x}{0.8-0.5}, & 0.5 < x \le 0.8 \\ 0 & 0.8 < x \le 1 \end{cases} \quad (7)$$

$$\mu_{k=good}(x) = \begin{cases} 0, & 0 \le x \le 0.4 \\ \dfrac{x-0.4}{0.7-0.4}, & 0.4 < x \le 0.7 \\ \dfrac{1-x}{1-0.7}, & 0.7 < x \le 1 \end{cases}$$

$$\mu_{k=very\_good}(x) = \begin{cases} 0, & 0 \le x \le 0.5 \\ \dfrac{x-0.5}{0.95-0.5}, & 0.5 < x \le 0.95 \\ 1, & 0.95 < x \le 1 \end{cases}$$

$$\mu_{a=high}(x)=\begin{Bmatrix}1, & 0\le x\le 0.1\\ \dfrac{0.75-x}{0.75-0.1}, & 0.1<x\le 0.75\\ 0, & 0.75<x\le 1\end{Bmatrix}$$

$$\mu_{a=average}(x)=\begin{Bmatrix}0, & 0\le x\le 0.1\\ \dfrac{x-0.1}{0.5-0.1}, & 0.1<x\le 0.5\\ \dfrac{0.9-x}{0.9-0.5}, & 0.5<x\le 0.9\\ 0 & 0.9<x\le 1\end{Bmatrix} \quad (8)$$

$$\mu_{a=poor}(x)=\begin{Bmatrix}0, & 0\le x\le 0.25\\ \dfrac{x-0.25}{0.25-0.9}, & 0.25<x\le 0.9\\ 1, & 0.9<x\le 1\end{Bmatrix}$$

$$\mu_{t=enough}(x)=\begin{Bmatrix}1, & 0\le x\le 0.1\\ \dfrac{0.75-x}{0.75-0.1}, & 0.1<x\le 0.75\\ 0, & 0.75<x\le 1\end{Bmatrix}$$

$$\mu_{t=average}(x)=\begin{Bmatrix}\dfrac{x-0.1}{0.5-0.1} & 0\le x\le 0.5\\ \dfrac{0.9-x}{0.9-0.5}, & 0.5<x\le 1\end{Bmatrix} \quad (9)$$

$$\mu_{t=not\_enough}(x)=\begin{Bmatrix}0, & 0\le x\le 0.25\\ \dfrac{x-0.25}{0.25-0.9}, & 0.25<x\le 0.9\\ 1, & 0.9<x\le 1\end{Bmatrix}$$

For the same answer values (customer input), the membership functions change, based on the answering order. If the answer to the question is the first one, the membership functions taper (10), what results in a more unique response.

$$\begin{aligned}
&\mu_{k=very_poor}{}^{1st}(x)=\left[\mu_{k=very_poor}(x)\right]^2\\
&\mu_{k=poor}{}^{1st}(x)=\left[\mu_{k=poor}(x)\right]^2\\
&\mu_{k=average}{}^{1st}(x)=\left[\mu_{k=average}(x)\right]^2\\
&\mu_{k=good}{}^{1st}(x)=\left[\mu_{k=good}(x)\right]^2\\
&\mu_{k=very_good}{}^{1st}(x)=\left[\mu_{k=very_good}\right]^2
\end{aligned} \quad (10)$$

$$\mu_{a=high}{}^{1st}(x) = \left[\mu_{a=high}(x)\right]^2$$
$$\mu_{a=average}{}^{1st}(x) = \left[\mu_{a=average}(x)\right]^2$$
$$\mu_{a=poor}{}^{1st}(x) = \left[\mu_{a=poor}(x)\right]^2$$

$$\mu_{t=enough}{}^{1st}(x) = \left[\mu_{t=enough}(x)\right]^2$$
$$\mu_{t=average}{}^{1st}(x) = \left[\mu_{t=average}(x)\right]^2$$
$$\mu_{t=not\_enough}{}^{1st}(x) = \left[\mu_{t=not\_enough}(x)\right]^2$$

If the answer to the question is the last one, the membership functions expand, what results in a more vague response.

$$\mu_{k=very\_poor}{}^{1st}(x) = \left[\mu_{k=very\_poor}(x)\right]^{0.9}$$
$$\mu_{k=poor}{}^{1st}(x) = \left[\mu_{k=poor}(x)\right]^{0.75}$$
$$\mu_{k=average}{}^{1st}(x) = \left[\mu_{k=average}(x)\right]^{0.25}$$
$$\mu_{k=good}{}^{1st}(x) = \left[\mu_{k=good}(x)\right]^{0.75}$$
$$\mu_{k=very\_good}{}^{1st}(x) = \left[\mu_{k=very\_good}\right]^{0.9}$$
$$\mu_{a=high}{}^{last}(x) = \left[\mu_{a=high}(x)\right]^{0.25} \quad (11)$$
$$\mu_{a=average}{}^{last}(x) = \left[\mu_{a=average}(x)\right]^{0.75}$$
$$\mu_{a=poor}{}^{last}(x) = \left[\mu_{a=poor}(x)\right]^{0.25}$$
$$\mu_{t=enough}{}^{last}(x) = \left[\mu_{t=enough}(x)\right]^{0.25}$$
$$\mu_{t=average}{}^{last}(x) = \left[\mu_{t=average}(x)\right]^{0.75}$$
$$\mu_{t=not\_enough}{}^{last}(x) = \left[\mu_{t=not\_enough}(x)\right]^{0.25}$$

As an example, for the same customer input (answer to the first question) of 0.65, but for different answering order, the membership functions are different, i.e. the values of the membership functions are also different, which is shown in Fig. 11.
The fuzzy output from the system, i.e. the decision is made in a manner that 45 if-then rules are defined. The rules are designed to produce three different outputs (o): "dummy", intermediate and professional. The membership functions in Fig. 12 are triangular or trapezoidal (12)

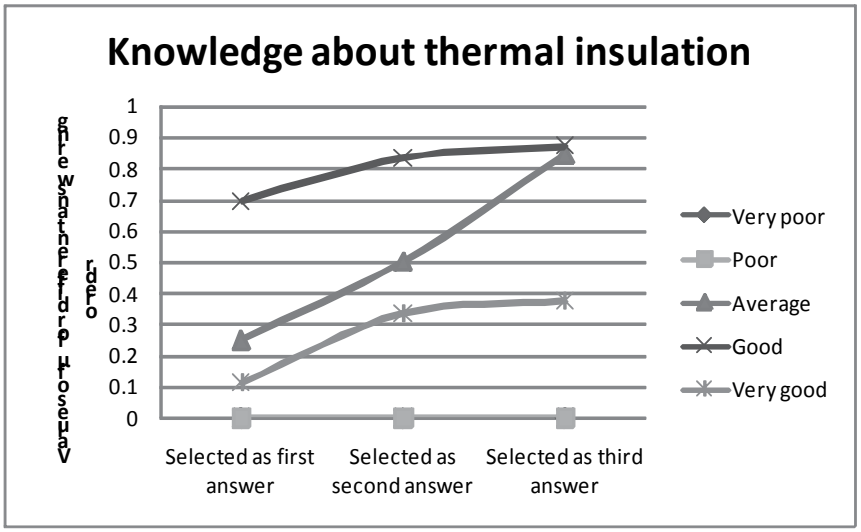


Fig. 11. Different values of the membership functions for different answering order

$$\mu_{o=dummy}(x)=\begin{Bmatrix}1, & 0\le x\le\alpha_0\\ \dfrac{\beta_0-x}{\beta_0-\alpha_0}, & \alpha_0<x\le\beta_0\\ 0, & \beta_0<x\le 1\end{Bmatrix}, \text{ where } \begin{matrix}\alpha_0=0.2\\ \beta_0=0.5\end{matrix} \text{ are the initial values}$$

$$\mu_{o=\text{int}ermediate}(x)=\begin{Bmatrix}0, & 0\le x\le\chi_0\\ \dfrac{x-\chi_0}{\delta_0-\chi_0}, & \chi_0<x\le\delta_0\\ \dfrac{\varepsilon_0-x}{\varepsilon_0-\delta_0}, & \delta_0<x\le\varepsilon_0\\ 0 & \varepsilon_0<x\le 1\end{Bmatrix}, \text{ where } \begin{matrix}\chi_0=0.3\\ \delta_0=0.5\\ \varepsilon_0=0.7\end{matrix} \text{ are the initial values} \quad (12)$$

$$\mu_{o=professional}(x)=\begin{Bmatrix}0, & 0\le x\le\phi_0\\ \dfrac{x-\phi_0}{\phi_0-\varphi_0}, & \phi_0<x\le\varphi_0\\ 1, & \varphi_0<x\le 1\end{Bmatrix}, \text{ where } \begin{matrix}\phi_0=0.5\\ \varphi_0=0.8\end{matrix} \text{ are the initial values.}$$

After the evaluation of if-then rules, an aggregated output is generated. Changes in input membership functions influence the customer profile configuration. For the same answers, but for a different answering order, the configured customer profile can be different.
The next example shows that for the following input data:

- 1st answer - customer input for knowledge about thermal insulation is 0.65;
- 2nd answer - customer input for accuracy of the configuration results is 0.8;
- 3rd answer - customer input for time for the configuration process is 0.5,

after defuzzification by the "Center of gravity method", the crisp output is 0.387 - and is interpreted as an “Intermediate customer”.

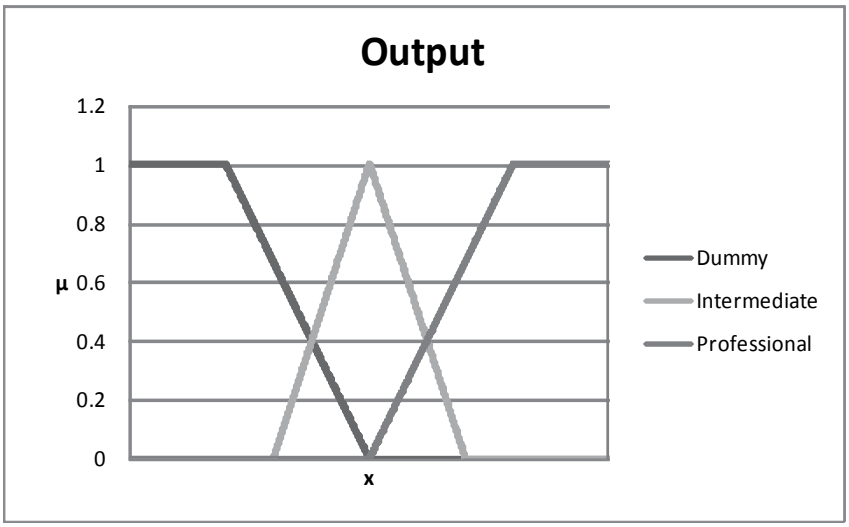


Fig. 12. Membership functions for output

For the following input data:

- 1st answer - customer input for accuracy of the configuration results is 0.8;
- 2nd answer - customer input for knowledge about thermal insulation is 0.65;
- 3rd answer - customer input for time for the configuration process is 0.5,

after defuzzification by the same method, the crisp output is 0.369 - and is interpreted as a "Dummy customer".

Based on the previous example, one can conclude that for the same input data, but for a different answering order, different customer profiles can be configured.

After the configuration task is finished, a feedback is generated. The customer is asked to answer a set of three questions:

- Are you satisfied with the complexity of the configurator? (c);
- Is the result satisfactory? (s);
- Are you satisfied with the time spent for the configuration process? (i).

The answers can range from "The configurator is too complex" (where the value of the answer is -0.5) to "The configurator is too easy " (where the value of the answer is 0.5) for the first question; from "The results should be more detailed and precise " (where the value of the answer is -0.5) to "The results are too detailed" (where the value of the answer is 0.5) for the second question; and from "I could have spent more time for the configuration process'' (where the value of the answer is -0.5) to "The configuration process was too long " (where the value of the answer is 0.5) for the third question. Initially, all the answers are set to the value of 0, which means that the customer is satisfied with the configuration process.

Based on the answers to questions, the input values for k, a, t are modified to k_{new}, a_{new}, t_{new} (13).

$$\begin{aligned} k_{new} &= k + \frac{c}{2} \\ a_{new} &= a + \frac{s}{2}, \\ t_{new} &= t + \frac{i}{2} \end{aligned} \qquad \begin{aligned} &0 \le k_{new} \le 1 \\ &0 \le a_{new} \le 1 \\ &0 \le t_{new} \le 1 \end{aligned} \tag{13}$$

This is the input for a new fuzzy output from the system, i.e. a new decision. This new output (o_{new}) takes into consideration whether a customer is satisfied with a configured customer profile. Based on the difference between an original and a new output, the membership functions for o_{i+1}, where o_{i+1} is the output in the future, are shifted left or right to better articulate the customers' preferences in the future. The amount of shifting is calculated in the following manner (6) as it was discussed before. The shifted membership functions for o are (14), with the following corrections (15).

$$\mu_{o=dummy}^{\,i+1}(x)=\left\{\begin{array}{cc} 1, & 0\le x\le \alpha_{i+1}=(\alpha_i+sa) \\ \dfrac{\beta_{i+1}-x}{\beta_{i+1}-\alpha_{i+1}}, & \alpha_{i+1}=(\alpha_i+sa)<x\le \beta_{i+1}=(\beta_i+sa) \\ 0, & \beta_{i+1}=(\beta_i+sa)<x\le 1 \end{array}\right\}$$

$$\mu_{o=\text{int}ermediate}^{\,i+1}(x)=\left\{\begin{array}{cc} 0, & 0\le x\le \chi_{i+1}=(\chi_i+sa) \\ \dfrac{x-\chi_{i+1}}{\delta_{i+1}-\chi_{i+1}}, & \chi_{i+1}=(\chi_i+sa)<x\le \delta_{i+1}=(\delta_i+sa) \\ \dfrac{\varepsilon_{i+1}-x}{\varepsilon_{i+1}-\delta_{i+1}}, & \delta_{i+1}=(\delta_i+sa)<x\le \varepsilon_{i+1}=(\varepsilon_i+sa) \\ 0 & \varepsilon_{i+1}=(\varepsilon_i+sa)<x\le 1 \end{array}\right\} \quad (14)$$

$$\mu_{o=professional}^{\,i+1}(x)=\left\{\begin{array}{cc} 0, & 0\le x\le \phi_{i+1}=(\phi_i+sa) \\ \dfrac{x-\phi_{i+1}}{\phi_{i+1}-\varphi_{i+1}}, & \phi_{i+1}=(\phi_i+sa)<x\le \varphi_{i+1}=(\varphi_i+sa) \\ 1, & \varphi_{i+1}=(\varphi_i+sa)<x\le 1 \end{array}\right\}$$

$$\begin{array}{l} if\ \alpha_{i+1}<0.05 then \alpha_{i+1}=0.05; if\ \alpha_{i+1}>0.35 then \alpha_{i+1}=0.35 \\ if\ \beta_{i+1}<0.35 then \beta_{i+1}=0.35; if\ \beta_{i+1}>0.65 then \beta_{i+1}=0.65 \\ if\ \chi_{i+1}<0.15 then \chi_{i+1}=0.15; if\ \chi_{i+1}>0.45 then \chi_{i+1}=0.45 \\ if\ \delta_{i+1}<0.35 then \delta_{i+1}=0.35; if\ \delta_{i+1}>0.65 then \delta_{i+1}=0.65 \\ if\ \varepsilon_{i+1}<0.55 then \varepsilon_{i+1}=0.55; if\ \varepsilon_{i+1}>0.85 then \varepsilon_{i+1}=0.85 \\ if\ \phi_{i+1}<0.35 then \varphi_{i+1}=0.35; if\ \phi_{i+1}>0.65 then \phi_{i+1}=0.65 \\ if\ \varphi_{i+1}<0.65 then \varphi_{i+1}=0.65; if\ \varphi_{i+1}>0.95 then \varphi_{i+1}=0.95 \end{array} \quad (15)$$

3.1 First results

The developed configurator has been tested configuring five existing buildings. The insulation is configured and the results are calculated for each customer profile. Heat loss is calculated, for input temperatures that are shown in Fig. 13.

Heat losses without insulation and with the proposed insulation, for different customer profiles are shown in Fig. 14 and Fig. 15, respectively.

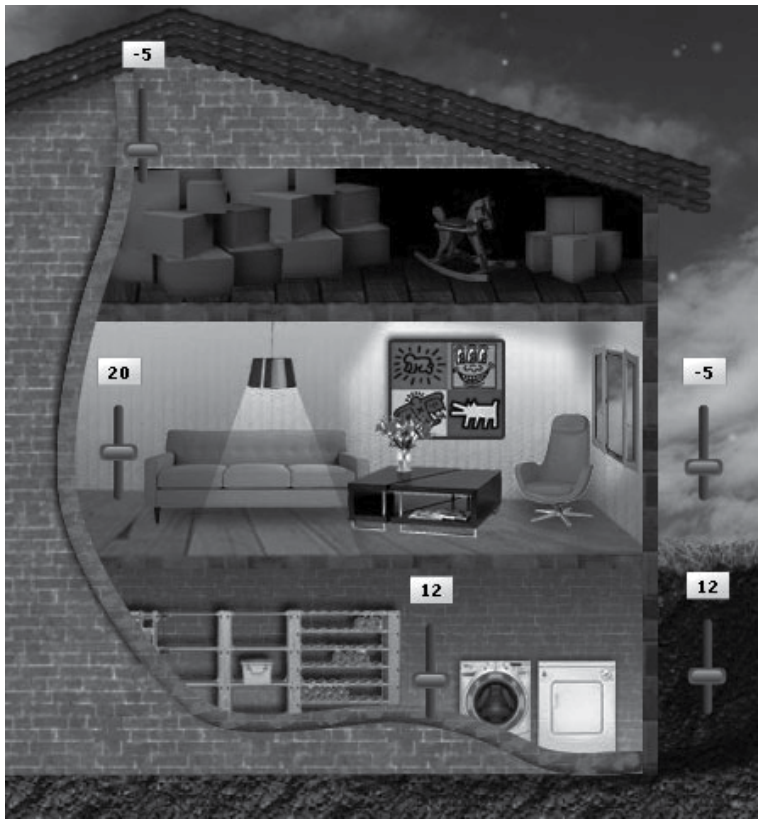


Fig. 13. Input temperatures

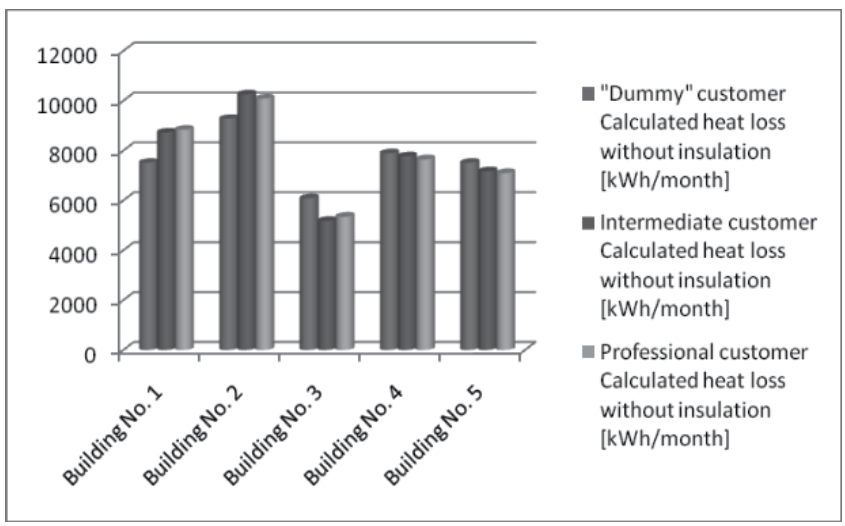


Fig. 14. Calculated heat loss without insulation

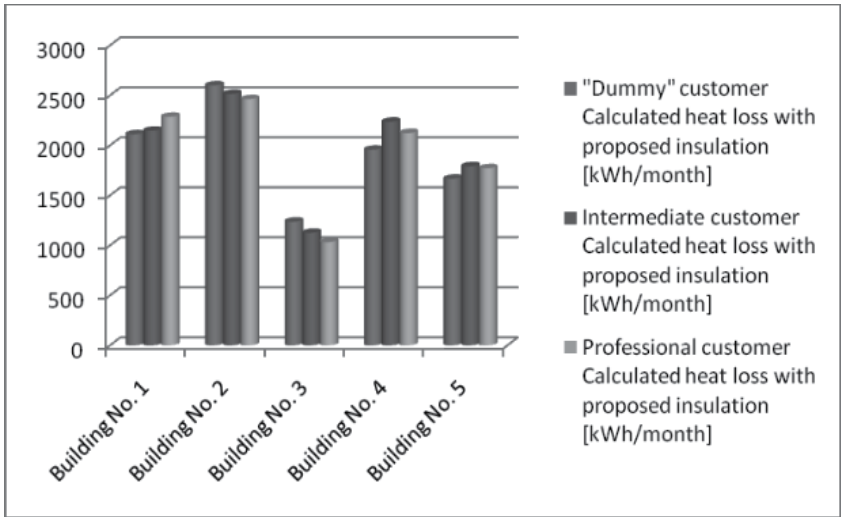


Fig. 15. Calculated heat loss with proposed insulation

Relative deviations of calculated heat losses without insulation and with the proposed insulation, for different customer profiles compared to detailed calculations are shown in Fig. 16 and Fig. 17, respectively.

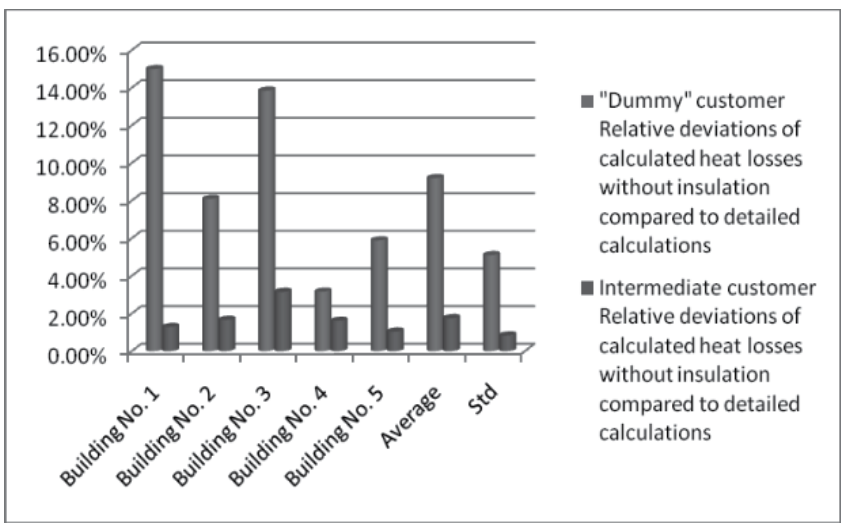


Fig. 16. Relative Deviations from detailed calculation without insulation

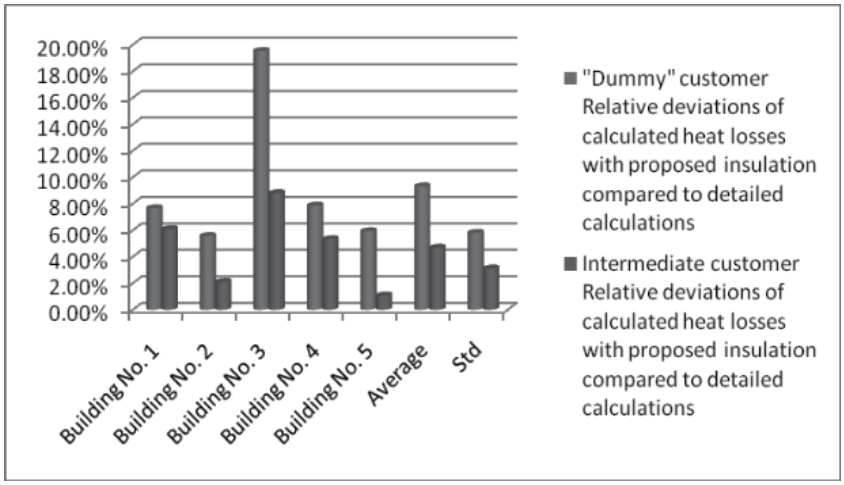


Fig. 17. Relative Deviations from detailed calculation with proposed insulation

4. Conclusion

The fact that in modern economy traditional product development is changed and moved towards a two-stage model, the first, the realm of company/designer establishing the solution space and the second, that of the customer as co-designer, fundamentally changes the role of the customer from the consumer of a product, to a partner in a process of adding value. This alteration of traditional product development through the involvement of the customer into the configuration of the final product faces some obvious problems. The fundamental challenge is to avoid the abortion of the configuration process by the customer.
The presented problem is solved by a proposed methodology for adaptive involvement of customers as co-creators in mass customization of products and services. The developed methodology identifies different customer profiles that suit each individual customer's needs and limitations.
The developed methodology is tested on a product configurator for thermal insulation of buildings. First results show that average deviation from the exact calculations for the "dummy" customer range from approximately 9.19% for calculations without thermal insulation to 9.31% for calculations with thermal insulation. Average deviation for the intermediate customer ranges from approximately 1.74% for calculations without thermal insulation to 4.68% for calculations with thermal insulation. Based on these results one can conclude that different customer profiles give different results, but that the differences could be accepted if the nature of the research field is taken into consideration.
The configuration process in the case of the "dummy" customer lasts about 3-4 minutes, for the intermediate customer the required time is about 5-10 minutes, and for the professional customer it takes more. The final solution is given in understandable form, which can be directly used for ordering. These results show that different customer profiles could be necessary for successful completion of the configuration process.
Experiences from retailers suggest that the idea of insulating a building is becoming more appealing and acceptable for the customers, when presented using the configurator, while end users suggest that there is further need to make the configurator more interesting.

The results and the gained experiences point towards several future research directions:

- Making the user interface more interesting by using as many visual and interactive elements as possible with real time multimedia help;
- Definition of rules for taking into account the accepted solutions by previous customers of certain profile and their incorporation into configurator;
- Development of an intelligent decision making algorithm that takes into consideration the general, specific and contextual information about customers during the customer profile generation as well as during the configuration process.

5. References

Berger, C. & Piller, F. (2003). Customers as Co-Designers, *IEE Manufacturing Engineer*, Vol. 82, No. 4, 42-46

Blecker, T. & Abdelkafi, N. (2006). Mass Customization: State-of-the-Art and Challenges. In *Mass customization: challenges and solutions*, Vol. 87, 1-25, Springer, New York

Bojadziev, G. & Bojadziev, M. (2007). *Fuzzy logic for business, finance, and management*, World Scientific Publishing, Singapore, ISBN 981-270-649-6

Chen, C. (2009). Human-centered product design and development. *Advanced Engineering Informatics*, Vol. 23, No. 2, 140-141

Chong, Y. T., Chen, C. & Leong, K. F. (2009). Human-centric product conceptualization using a design space framework, *Advanced Engineering Informatics*, Vol. 23, No. 2, April 2009, 149-156, ISSN 1474-0346

Čović, Z., Fürstner, I., Anišić, Z. & Freund, R. (2009). Web Based Intelligent Product Configurator for Thermal Insulation and Decoration of Buildings, *Proceedings of 7th International Symposium on Intelligent Systems and Informatics*, ISBN 978-1-4244-5349-8, Subotica, Serbia, Sept. 2009, (CD)

Engelbrektsson, P. & Soderman, M. (2004). The use and perception of methods and product representations in product development: a survey of Swedish industry. *Journal of Engineering Design*, Vol. 15, No. 2, April 2004, 141-154, ISSN 0954-4828

Forza, C. & Salvador, F. (2007). *Product Information Management for Mass Customization*, Hampshire: Palgrave Macmillan

Franke, N. & Piller F. (2003). Key Research Issues in User Interaction with Configuration toolkits in a Mass Customization System, *International Journal of Technology Management*, Vol. 26, No. 5/6, 578-599

Fürstner, I. & Anišić, Z. (2009a). Intelligent Product Configurator – The New Approcach in Thermo Insulation of Buildings, *Journal of Engineering Annals of the Faculty of Engineering Hunedoara*, Vol. 7, No. 2, 165-170, ISSN 1584-2665

Fürstner, I. & Anišić, Z. (2009b). Self-Adaptive Product Configurator for Thermal Insulation. *Proc. Tenth Int. Symposium of Hungarian Res. on Comp. Intelligence and Informatics*, pp. 669-680

Fürstner, I. & Anišić, Z. (2009c). Adaptive Product Configuration for Thermal Insulation of Buildings. *Proc. Twentieth Int. DAAAM Symposium "Intelligent Manufacturing & Automation: Theory, Practice & Education"*, pp. 1037-1038, Vienna, Austria: DAAAM International Vienna

Galbraith, J. R. (2005). *Designing the Customer-Centric Organization*, Jossey-Bass, ISBN 0-7879-7919-8, San Francisco

Gero, J. S. & Kannengiesser, U. (2004). The situated function-behaviour-structure framework, *Design Studies*, Vol. 25, No. 4, 373-391

Hansen, T., Scheer, C. & Loos, P. (2003). Product Configurators in Electronic Commerce – Extension of the Configurator Concept - Towards Customer Recommendation, *Proceedings of the 2nd Interdisciplinary World Congress on Mass Customization and Personalization (MCP)*, Technische Universitaet Muenchen Munich

Hanss, M. (2005). *Applied Fuzzy Arithmetic, An Introduction with Engineering Applications*, Springer, ISBN 3-540-24201-5, Berlin

Koch, M. & Moeslein, K. (2003). User Representation in eCommerce and Collaboration Applications, *Proceedings of the 16th Bled eCommerce Conference eTransformation*, pp. 649-661, Bled, Slovenia, June 2003

Kumiawan, S., Tseng, M. & So, R. (2003). Consumer Decision-Making Process in Mass Customization, *Proc. Second Int. World Cong. on Mass Customization and Personalization*

Leckner, T. & Lacher, M. (2003). Simplifying configuration through customer oriented product models, *Proceedings of the International conference on engineering design ICED 03*, August 2003, Stockholm, Sweden

Levy, A. & Weld, D. (2000). Intelligent Internet Systems. *Artificial Intelligence*, Vol. 118, No. 1-2, April 2000, 1-14, ISSN 0004-3702

Maravić, Č. P., Čisar, P & Pinter, R. (2009).True/false Questions Analysis Using Computerized Certainty-based Marking tests, *Proceedings of the 7th International Symposium on Intelligent Systems and Informatics SISY*, Subotica, Serbia, September 2009., Proceedings CD ROM, ISBN 978-1-4244-5349-8

Reichwald, R., Seifert, S., Walcher, D. & Piller, F. (2004). Customers as part of value webs: Towards a framework for webbed customer innovation tools, *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Hawaii

Schubert, P. & Koch, M. (2002). The Power of Personalization: Customer Collaboration and Virtual Communities, *Proceedings of the Eighth Americas Conference on Information Systems*, pp. 1953-1965

Zimmermann, H. J. (1998). *Fuzzy set theory – and its applications*, Kluwer-Nijhoff Publishing, Boston

11

The Market for Nanotechnology Applications and Its Managerial Implications: An Empirical Investigation in the Italian Landscape

Lucio Cassia and Alfredo De Massis
University of Bergamo
Italy

1. Introduction

In the last 7 to 8 years attention towards nanotechnology and its applications has greatly increased, both in academic and industrial environments. Its potential was recognised earlier in the 1959 speech by Nobel Laureate Richard Feynman, 'There's Plenty of Room at the Bottom', at the annual meeting of the American Physical Society at the California Institute of Technology (Caltech). The great expectations that surround nanotechnology are witnessed by the growth of public funds devoted to research into the field. In 2004, these funds reached the amount of approximately 4.6 billion US Dollars worldwide, with an increase of more than 700% with respect to 1997 (www.luxresearchinc.com, 2004). Several European countries started national nanotechnology research programs between the mid-1980's and mid-1990's. Here the overall investment of around 200 million Euro in 1997 has risen to approximately 1,000 million Euro in 2003 (www.nanoinvestors.com, 2004). Moreover, nanotechnology was identified as a main priority area in the 6th Framework Programme (~250 M€/year). Italy, with over 1,200 people directly involved and 70 million Euro of R&D funding, is shaping up to be a major world player in nanotechnology. Here significant nanotech investment have been realised, both in private and public sectors, and dedicated research centres have been established (www.nanotec.it, 2004).

Today, a market for nanotechnology's applications is therefore emerging. Even if still limited in dimensions, its value is approximately of 50-60 million U.S Dollars (European Commission, 2004a, b), but is thought by analysts to be capable of rising to hundreds of billion sale volumes by 2010, and exceeding one trillion after (www.nanobusiness.org, 2000). However, the novelty of the nanotech market, the pervasiveness of nanotechnologies and the complexity of their applications, the lack of common definitions, concepts and market boundaries, make it very difficult to clearly distinguish between nanotech and non-nanotech companies.

In this paper, building on previous research on nanotechnology firms classification (Chiesa & De Massis, 2006), we label as 'nanotech' a company that: (i) carries out R&D activities on nanotechnology (devoting human, technological and financial resources to its study); (ii) can be easily arranged in one or more of the four "nanotechnology categories" (nanomaterials, nanotools, nanostructures, and nanoprocesses) considered in the framework developed by Chiesa et al. (2005a).

The objectives of this paper are twofold:

- to offer a picture of the overall Italian market for nanotech applications, i.e. to identify the main Italian players operating in the nanotechnology and to collect the data and information needed in order to understand the general characteristics if this market (companies' dimension, geographical location, industry, main activities, and nanotechnology category(ies) in which they operate); and,
- to provide, through the analysis of some relevant variables, a description of the management and organisation of the players operating in such a market, drawing out the emerging business models and understanding the determinants of these configurations in light of the specific characteristics of the different players and the nanotechnology categories in which they operate.

Given these objectives, this paper deliberately does not consider either the academic/public research organisations or the private research centres involved in nanotechnology, but is focused solely on Italian private industrial nanotech companies which operate in order to commercialise nanotech applications.

The paper is structured as follows: Section 2 describes the research methodology; Section 3 illustrates the empirical study and discusses the empirical results; finally, Section 4 draws some conclusions, considers the implications for academics and practitioners, and suggests future directions of research.

2. Research methodology

As previously pointed out, the paper has two main objectives:

The first objective is to describe the Italian market for nanotech applications, i.e. to identify its main players and collect some general information useful for understanding the market as a whole, such as: (i) the companies' dimensions; (ii) the localisation; (iii) the main activities undertaken. As far as this objective is concerned, an extensive analysis was conducted and focused on the following topics: (i) diffusion of nanotech firms; (ii) aspects related with the localisation of these companies; (iii) type of activities conducted and nanotechnology categories where they work. The main problems in organising the extensive analysis were related to the novelty of the nanotech market, the pervasiveness of nanotechnologies and the complexity of their applications, the lack of common definitions, concepts and market boundaries. Considering these difficulties, it was decided to base the research on the web, searching for the websites of companies that develop, produce and/or use nanotech products or processes. The main limitations associated with the use of Internet as research tool are related to the generally low validity of the information collected (Hewson et al., 2003). This is mainly due to the: (i) subjectivity introduced by the researcher which, on the basis of his personal characteristics, is led to select some information rather than others; (ii) lack of controls over the material published on the web. However, such limitations can be easily overcome by considering that: (i) the extensive analysis needs general and objective information, mostly descriptive in nature and publicly available, which are usually well suited to accurate and precise publication on the web; (ii) the great attention from information-providers directed towards those few Italian companies operating in nanotechnology has led to a wide diffusion of information on the web about such companies, thus giving the opportunity to validate their reliability through the comparison of different online sources of information. The data collected were then integrated with reports drawn from various sources (www.luxresearchinc.com, 2004; www.nanoinvestornews.com, 2004;

European Commission, 2004a; www.investinitaly.com, 2004; www.nanovip.com, 2004; www.nanotec.it, 2004; Nanotech/It, 2004; Caravita, 2004a, b).
The second objective of the paper is to provide a description of the management and organisational practices adopted by Italian nanotech players, investigating the emerging business models. According to this objective, it was decided to conduct an intensive analysis based on a multiple case study aimed at in-depth analysing the previously identified sample of Italian nanotech companies. It is worth mentioning that multiple case study research appears here to fit with the objective of the analysis (Yin, 2003; Eisenhardt, 1989) since the investigated phenomenon is very complex and hardly distinguishable from its context (i.e. the competitive environment in which firms operate), and the investigation of the business models requires to explore processes and activities and answering "how" and "why" questions, providing explanations rather than statistical information. Only 14 of the 19 identified nanotech companies (74% of the population) gave their availability to cooperate in the research.
Two semi-structured telephone or personal interviews were conducted with all the top – or research – managers of the 14 companies between the end of 2004 and the first half of 2005, through a questionnaire specifically designed in order to analyse the business models of the companies. Given that many authors have offered definitions of the term 'business model' (see Chesbrough & Rosenbloom, 2002), but no generally accepted definition of a business model has emerged to date (Shafer et al., 2005), the business model as used in this study is the set of answers to the following five questions (Hellström & Sjölander, 2005): (i) What is the offer to the customer? (ii) Who is the customer? (iii) What is the value thereby created for the customer? (iv) How can the firm appropriate a sufficient share of that value? (v) How is that offer conveyed to the customer? Have been therefore explored variables related to: (i) the nanotech output and the characteristics of the customers to which it is offered (type of output, method of use of nanotechnology, type of served market, destination of the output in terms of industries of origin of the main customers); (ii) the mechanism adopted by the company for entering the nanotech business and the degree of correlation between the core and the nanotech business; (iii) the way the company manages the sale of nanotech output (marketing approach, model of involvement of the client company, degree of standardisation/ customisation of the provided output, phase(s) of interaction with the client company, commercial relationship management model); (iv) the structure adopted in order to organise the technical personnel and the relating activities; (v) the intellectual property management model and the strategies used for solving the appropriation problems.
The interviews were integrated with internal and public documentations (newspaper articles, websites, balance sheets) and archival records (list of products, organisational charts, service records), thus allowing the triangulation of empirical evidence (Patton, 1987).

3. The empirical study

The empirical study has been divided into two distinct parts:

- the extensive analysis;
- the intensive analysis.

3.1 The extensive analysis: basic results

According to the definition of 'nanotech company' adopted in this study, 19 firms were identified in the Italian landscape. They are listed in Table 1, together with their location and industry of origin.

Company	Location	Industry
APE Researech	Basovizza (TS)	Instrumentation
CRF (Centro Ricerche Fiat)	Torino	Automotive
CSM	Roma	Materials
EniTecnologie	San Donato Milanese (MI)	Energy
Geal	Agliana (PT)	Chemicals
Kedrion	Castelvecchio Pascoli-Barga (PT)	Pharmaceuticals
MBN	San Vendemiano (TV)	Materials
Microcoat	Sedriano (MI)	Coatings
Moma	Reggiolo (RE)	Coatings
Olivetti I-Jet	Arnad (AO)	Microelectronics
Organic Spintronics	Bologna	Materials
Pirelli Labs	Milano	Information Technology
Pometon	Maerne di Martellago (VE)	Materials
SAES Getters	Lainate (MI)	Instrumentation
Scriba Nanotecnologie	Bologna	Materials
SIAD	Bergamo	Chemicals
Sorin Biomedica Cardio	Saluggia (VC)	Biomedicine
STMicroelectronics	Agrate Brianza (MI)	Microelectronics
Tethis	Milano	Coatings

Table 1. Italian nanotech companies.

First, it is therefore possible to observe the low number of players compared to other countries (e.g. U.S., Germany, France and Japan), in confirmation of the embryonic of stage of the Italian market for nanotech applications.
The nanotech companies are mostly located in the northern part of the country (15 firms, corresponding to 84% of the sample); three of them (16% of the sample) are located in the central part of Italy. No companies have been found in the South or in the islands. The geographical distribution of nanotech companies is shown in Figure 1.
Interesting findings emerge when the industries in which the nanotech companies operate are analysed. These are summarized in Table 2.

Industry	**Number (%) of companies operating in the industry**
Materials	5 (26)
Coatings	3 (16)
Microelectronics & ICT	3 (16)
Chemicals	2 (11)
Instrumentation	2 (11)
Energy	1 (5)
Pharmaceuticals	1 (5)
Biomedicine	1 (5)
Automotive	1 (5)

Table 2. Number (and percentage) of Italian nanotech companies operating in different industries

Fig. 1. Geographical distribution of Italian nanotech companies.

It becomes clear that in Italy nanotechnologies are mainly applied in the development and production of improved materials, which, at the moment, seem to be the most promising nanotech applications. They are also being used in the engineering of enhanced coating treatments and for the realization of microelectronic and ICT devices. Chemicals and instrumentation are other important industrial sectors where nanotech applications are being implied.

In Table 3, the 19 Italian companies are classified into the four nanotechnology categories identified in the nanotechnology categorisation framework proposed by Chiesa et al. (2005a).

Nanotechnology categories			
Nanomaterials	**Nanotools**	**Nanostructures**	**Nanoprocesses**
CSM	APE Research	CRF	EniTecnologie
Geal		Kedrion	Microcoat
MBN		Olivetti I-Jet	Moma
Pometon		Pirelli Labs	Organic Spintronics
SIAD		Sorin Biomedica Cardio	Scriba Nanotecnologie
		STMicroelectronics	SAES Getters
			Tethis

Table 3. Distribution of Italian nanotech companies in the nanotechnology categories.

It is possible to note that:

- the majority of companies (7, corresponding to 37% of the sample) use nanotechnologies in order to carry out, for the client firm, specific nanotech processes (e.g. magnetic abrasive finishing, thin film deposition);

- 6 firms (32% of the sample) apply nanotech intermediate findings (e.g. nanotubes), instrumentation operating at the nanoscale (e.g. piezodriven nanopositioners) or nanotech processes (e.g. nanolithography) in order to produce and commercialise innovative devices classifiable as nanostructures, i.e. complex systems made of different parts matched together that perform different functions (e.g. the Nano-Electro-Mechanical-Systems (NEMS) fabricated by STMicroelectronics);
- 5 companies (26% of the sample) apply nanotechnologies in order to manufacture innovative materials (e.g. nanopowders);
- just one firm (A.P.E. Research) uses nanotechnologies in order to fabricate and sell to client companies high-precision tools operating at the nanoscale (e.g. atomic force microscopes).

Finally, considering the dimensions of the analysed companies, it is possible to identify:

- 7 big companies (STMicroelectronics, SIAD, SAES Getters, Centro Ricerche Fiat-CRF, Sorin Biomedica Cardio, EniTecnologie, Kedrion) that employee 70% of the overall personnel working in the field;
- 4 medium-sized enterprises (Olivetti I-Jet, CSM, Pometon, Pirelli Labs), with a number of employees between 200 and 344;
- 2 small companies (Geal, MBN), employing 32 and 35 persons respectively;
- 6 micro-firms, with less than 9 employees. They are start-ups (Microcoat, Moma) or academic spin-offs (Tethis, APE Research, Organic Spintronics, Scriba Nanotecnologie) recently founded (the oldest company is APE Research, founded in 1996).

The intensive analysis, the results of which are described in the next session, includes the study of the management and organisational practices adopted by Italian nanotech companies, and the investigation of the interrelations between such practices, the dimension of the companies, and the nanotechnology category in which they work.

3.2 The intensive analysis: basic results

The empirical evidence collected during the intensive analysis is schematically reported in Table 4.

The main results of the intensive analysis can be summarised as follows.

- Significantly, it is possible to identify four methods according to which the companies use nanotechnology to obtain the commercialisable output:
  - M1. the method adopted by nanotech companies developing intermediate findings (e.g. nanopowders) to be sold, licensed out or partnered with the client company, that carries out the remaining tasks of the development process and includes them into its innovative products. These companies provide a service of "Work In Progress (WIP) Innovation" (Chiesa et al., 2005b) and typically operate in the category of nanomaterials;
  - M2. the method adopted by nanotech companies carrying out, for the client firm, a particular process (e.g. thin film deposition) that requires excellent competencies in a specific scientific domain. These companies provide a service of "Process Activity" (Chiesa et al., 2005b), and usually operate in the nanoprocesses category.
  - M3. the method adopted by nanotech companies developing and selling instrumentation operating at the nanoscale (e.g. Scanning Probe Microscopes or software for real-time imaging) that is used by the client firm for supporting its innovation processes (typically, its basic and applied research activities). These

companies provide a service of "Technologies to develop technology" (Debackere, 1999) and typically operate in the category of nanotools;

Nanotechnology Categories																		
	Nanomaterials			Nanotools	Nanostructures					Nanoprocesses								
Variables	Goal	Pometon	SIAD	APE Research	Olivetti I-Jet	Kedrion	SorinBiomedica Cardio	STMicroelectroni cs	Mossa	SorinNanotecnologie	Tethis	Organic Spintronics	SAES Getters	Eni Tecnologie
Company's dimension	Small	Medium	Big	Microscopic	Big	Big	Big	Big	Microscopic	Microscopic	Microscopic	Microscopic	Big	Big
Type of nanotech output	Nanotech intermediate finding	Nanotech intermediate finding	Nanotech intermediate finding	Tecnology	Product including nanotech	Product including nanotech	Product including nanotech	Product including nanotech	Process service	Process service	Process service	Process service	Process service	Process service
Method of use of nanotechnology	M1	M1	M1	M3	M4	M4	M4	M4	M2	M2	M2	M2	M2	M2
Served market	Existing	Existing / New to the company	Existing / New to company	*	Existing / New to the company	Existing / New to the company	Existing	Existing / New to the company	*	*	*	*	Existing / New to the company	Existing
Destination of nanotech output	Different sectors	Different sectors	Different sectors	Different sectors	Different sectors	Different sectors	Single sector	Different sectors	Different sectors	Different sectors	Different sectors	Different sectors	Different sectors	Different sectors
Nanotech business entry mechanism	Internal development	Internal development	Internal development	Academic spin-off	Internal development	Internal development	Internal development	Internal development	Start-up	Academic spin-off	Academic spin-off	Academic spin-off	Internal development	Internal development
Degree of correlation between core and nanotech business	High	High	High	*	High	Medium	High	Medium	*	*	*	*	High	High
Marketing approach	Structured	Structured	Not structured	Structured	Structured	Not structured	Structured	Structured	Not structured	Semi-structured	Semi-structured	Semi-structured	Structured	Structured
Model of involvement of the client company	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct	Direct
Standardization/customization of nanotech output	Standardized	Standardized	Standardized	Customized	Standardized	Standardized	Standardized	Standardized	Customized	Customized	Customized	Customized	Customized	Customized
Phase of interaction ith the client company	Delivery	Delivery	Delivery	R&D	Delivery	Delivery	Delivery	Delivery (occasionally R&D)	R&D	R&D	R&D	R&D	R&D	R&D
Commercial relationship management model	Not structured	Not structured	Not structured	Structured	Not structured	Not structured	Not structured	Not structured (occasionally structured)	Semi- structured	Not structured	Semi-structured	Semi-structured	Structured	Structured
Organizational structure	Input oriented in Output oriented	Output oriented	Input oriented in Output oriented	Weak matrix	Weak matrix	Input oriented for activity	Input oriented for activity	Input oriented for activities	Input oriented for scient. discip.	Input or/ weak matrix	Input oriented for activity	Input or / weak matrix	Weak matrix	Weak matrix
Intellectual property management form	Industrial secret	Industrial secret	Proprietary brands, industrial secret	None	Patents, Propr. brand, Copyright, Ind. secret	Patents, Propr. brands, Ind. secret	Patents, Propr. brands, Ind. secret	Patents, Ind. secret	Industrial secret	Patents, Ind. secret	Patents, Ind. secret	Patents, Ind. secret	Patents, Propr. brands, Ind. secret	Patents, Propr. brand
Payment system for solving the client's firm appropriation problems	Spot payment	Spot payment	Spot payment	Spot payment + royalty	Spot payment + royalty	Spot payment	Spot payment	Spot payment	Spot payment	Spot payment	Spot payment + royalty	Spot payment	Spot payment	Spot payment + royalty

Table 4. The empirical evidence emerged from the intensive analysis (the symbol '*' means that for start-ups and academic spin-offs the value of the variable is senseless).

M4. the method adopted by those nanotech companies internally developing, acquiring or licensing in from other companies nanotechnological processes, tools, or intermediate findings to be used for innovating their processes or products; it is typical of companies operating in the nanotechnology category of nanostructures.

- All companies use nanotechnologies to serve an existing market, even if in some cases the nanotech output opens a new market. This shows that the Italian nanotech companies are more inclined to exploit nanotechnology in order to strengthen the position in their core business rather than to diversify into different businesses. This is proved by the basically high degree of correlation between the core and the nanotech business.
- The firms, with the only exception of Sorin Biomedica Cardio, address their nanotech output to companies belonging to different industrial sectors. Nanotechnology, in fact, has a great pervasiveness, that brings companies offering an output in this scientific domain to direct their offer to firms operating in various sectors of activities; this is necessary if they mean: (i) to fully exploit the potentiality of the technical knowledge they possess; (ii) to make economically bearable the huge investments in nanotech applications. In the case of Sorin Biomedica Cardio, the choice to direct the offer exclusively to the biomedical sector is dictated by the corporate strategy of Sorin Group, which is uniquely focused on the sector of biomedicine, and the peculiarity of its nanotech applications.
- All companies have entered the nanotech business through the mechanism of internal development (even the start-ups and spin-offs have internally developed their know-how on nanotechnologies). This unequivocal choice can be explained in the light of three main factors: (i) the high degree of correlation between the core and the nanotech business, which has pushed the firms to internally develop the nanotech competencies; (ii) the embryonic stage of the Italian nanotech market, which has made difficult to find external holders of nanotech competencies; (iii) the limited experience of Italy in corporate venturing activities, that has hindered the adoption of new business entry mechanisms different from internal development (e.g. venture capital investments).
- Almost all companies adopt a structured marketing approach based upon the dispatch of information packs to potential clients, the participation to professional fairs and the use of the website as a window on the firm's projects and services. The contact of new potential clients seems therefore to be a critical aspect for the analysed firms. The only cases of non-structured marketing approach are that of: (i) Kedrion and SIAD, since they are still far from the phase of commercialisation of nanotech applications (both the companies have in fact revealed that will likely adopt a structured marketing approach once arrived in the commercialisation phase); (ii) Moma (in this case, the extremely small dimension (four employees) of the firm prevents it from possessing the financial and human resources needed for undertaking structured marketing initiatives);
- All companies adopt a direct model of involvement of the client company, i.e. interact personally with the client firm's reference people. The high complexity of nanotechnology and the current scant knowledge of its applications by client firms make personal contact very important with the client and prevent nanotech players from adopting indirect models of client's involvement (e.g. online selling).
- Considering the other three dimensions of the sale management model, i.e. the commercial relationship management approach, the phase of interaction with the client company and the degree of standardization of the provided service, it could be stated that two alternative models seem to be applied by Italian nanotech firms in the management of the service sale. They are:

- Customized model, characterized by: (i) customized output; (ii) interaction with the client companies since the earliest phase of development; (iii) structured commercial relationship management model.
- Standardized model, characterized by: (i) standardized output; (ii) interaction with the client companies only in the phase of output delivery; (iii) not structured commercial relationship management model.

• No specially interesting findings seem to appear when the organisational structure of the companies, the intellectual property management model, and the strategies used for solving the appropriation problems are considered.

3.3 Emerging business models among italian nanotech companies

The collected empirical evidence suggests the existence of correlations between: (i) the nanotechnology category in which the companies work; (ii) their dimension; (iii) the adopted sale management model. It seems therefore that four main types of business model can be identified among the Italian nanotech players (Figure 2).

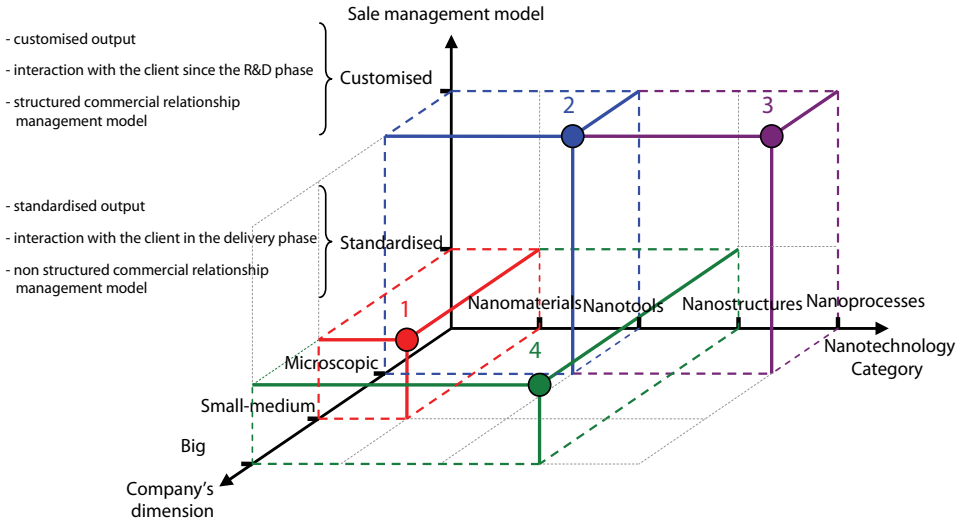


Fig. 2. The emerging business models among Italian nanotech companies.

1. The configuration generally adopted by small and medium sized enterprises operating in the nanotechnology category of nanomaterials with a standardised sale management model (Geal, Pometon). They are companies offering a service of "Work In Progress (WIP) innovation". The adoption of a standardised sale management model can be partially understood if it is considered that a "WIP innovation" is an intermediate finding that needs to be further developed by the client company and included in its final products. In order to allow that this intermediate finding is incorporated into the widest range of finished products, it must have general properties. The case of SIAD might seem an exception since it operates in the category of nanomaterials and adopts a standardised sale management model, but is a big company; anyway, the part that within the company is dedicated to the business of nanomaterials has a small dimension.
2. The configuration adopted by microscopic companies operating in the category of nanotools with a customised sale management model (A.P.E. Research). They are companies offering a service of "Technologies to develop technology" that is used by the

client firm for supporting its innovation processes. In order to understand why "technologies to develop technology" are sold according to a customized model, it is necessary to remember that nanotechnology has a highly interdisciplinary feature. Therefore, a nanotech-based tool (e.g. a Scanning Probe Microscope) can be useful in supporting the client firm's innovative activities only if it fits its particular requirements.

3. The typical configuration of microscopic firms operating in the category of nanoprocesses with a customised sale management model (Moma, Organic Spintronics, Scriba Nanotecnologie, Tethis). They are companies providing a service of "Process activity". The adoption of a customised sale management model in this case is reasonable if it is considered that "process activity" is a cluster of services that is necessarily highly customized; the actual necessities of the innovator, in fact, strictly influence the type of activities to be undertaken by the "outsourcer" nanotech company. The cases of SAES Getters and EniTecnologie might seem exceptions since they work in the category of nanoprocesses and adopt a customised sale management model, but have a big dimension; however, if the activities undertaken are carefully considered, it clearly emerges that the part that within the companies is dedicated to the nanotech business has a microscopic dimension.
4. The configuration adopted by big companies working in the category of nanostructures with a standardised sale management model (Olivetti I-Jet, Kedrion, Sorin Biomedica Cardio, STMicroelectronics). They are companies that apply nanotechnologies, internally developed or acquired from other companies, for innovating their products. They are therefore companies that sell finished products including nanotechnologies. The only firms operating in the category of nanostructures are big ones, and this can be partially explained by considering that nanostructures are complex systems (e.g. molecular logic circuits, organic LEDs) that require, in order to be fully developed and sold to the final customer, huge investments that can be afforded uniquely by big companies; a small company, in fact, seldom possesses the financial resources needed in order to finish the manufacture of a nanostructure to be sold to the client. Furthermore, the adoption of a customised sale management model can be explained in the light of the fact that nanostructures are finished systems capable of autonomously functioning that the customer can use without any need for modification and customisation.

The classification of the business models of the Italian nanotech companies in the four aforementioned configurations seems to be a viable instrument for further analysing nanotech companies from a managerial perspective. It has in fact proved to be capable of explaining significant differences among the players in terms of: (i) nanotechnology category where they work; (ii) dimension; (iii) sale management model.

Unluckily, at the moment, the number of nanotech companies in Italy is very low and it will be necessary to wait for the development and proliferation of new companies in order to verify if the identified business models represent a common trend of development in particular clusters of nanotech firms.

4. Conclusions and implications

Interest towards nanotechnology has largely grown in the last 7 to 8 years, and this scientific field has proved to be particularly promising for the development of an emerging and rapidly-growing market for nanotech applications. The paper offers a picture of the overall Italian market for nanotech applications and provides an insight into the management and organisation of the players operating in such a market.

In order to achieve this purpose, an empirical study was conducted. It was divided into two

distinct parts: (i) an extensive analysis, aimed at identifying the main players of the Italian market for nanotech applications and collecting some general information useful for understanding the market as a whole; and (ii) an intensive analysis, aimed at investigating the management and organisational practices adopted by Italian nanotech players. The collected empirical evidence has allowed the identification of some emerging business models adopted by Italian nanotech companies. An explanation of the determinants of these business configurations in light of the specific characteristics of the different players and the nanotechnology categories in which they operate has then been provided.

The results of the empirical study have significant implications both for academics and practitioners: First, it is believed that the present research can benefit researchers interested in the study of the nanotech market and, in general, the high-tech industries. They are encouraged to adopt the research methodology suggested in the paper in order to compare the major findings in the Italian landscape with the case of other countries and high-tech industries. This would help to extend the present research's results and advance the extant knowledge about the development trends and the emerging business models within growing high-tech industries.

Second, the paper suits as a background policy document for policy makers and investors in Italy. The high-tech industries are being paid increasing attention in the design of public policies (U.S. Government, 2006, 2003), and particularly the nanotech industry is considered critical for favouring the development of advanced economies (European Commission, 2004c). In this respect, the paper's achievements are useful since they suggest how to build a system of supporting initiatives that fits with the current evolution and characteristics of the Italian nanotech industry; these initiatives, in fact, are likely to be far more effective, in a specific stage of the industry's life cycle, if they aim at stimulating the diffusion of those firms that adopt the types of business models which are contextually capable of generating and delivering the highest value added.

Third, the paper provides corporate executives with useful insights into the role played by nanotechnology within a business context and the way the nanotech companies are currently managing and organising their business activity. In order to define the optimal business model for their companies, the executives of firms aimed at commercialising nanotech applications should first consider the current characteristics of the nanotech market and its stage of development, and then adapt the role they mean to play in the market to the figure that is likely to be precursory of the highest value.

The empirical study has made a step further in the investigation of the growing market for nanotech applications, offering some empirical evidence about the business models applied by companies operating in the nanotechnology domain. An interesting future direction of research consists in expanding the analysis here presented to other countries and high-tech industries (e.g. the biotech sector). This would give the opportunity to understand whether the major findings described in this paper prove valuable in other contexts and, if not, what the contextual variables (country-specific and/or industry-specific) might be that would explain the emerging differences. The ongoing research project means to shed light on this topic.

5. References

Caravita, G. (2004a) Nanotech, arriva la mappa italiana. *@lfa Il Sole-24 Ore*, 15 July.

Caravita, G. (2004b) Il nanotech cala un tris d'assi. *@lfa Il Sole-24 Ore*, 2 December.

Chesbrough, H. & Rosenbloom, R. S. (2002) The role of the business model in capturing value from innovation: evidence from Xerox Corporation's technology spin-off companies. *Industrial and Corporate Change*, 11, 3, 529-555.

Chiesa, V. & De Massis, A. (2006) *La nanoindustria: analisi dei principali player italiani nelle nanotecnologie.* Roma: Aracne Editrice.

Chiesa, V., De Massis, A. & Frattini, F. (2005a) Emerging business models in TSS: the case of nanotechnology in Italy. In Khalil T. (eds.) *Productivity Enhancement for Social Advance: The Role of Management of Technology – IAMOT 14th International Conference on Management of Technology Proceedings*, Vienna, Austria, 22-26 May.

Chiesa, V., Frattini, F., Lazzarotti, V. & Manzini, R. (2005b) Managing and organizing technical and scientific service companies: a conceptual framework and an empirical study. In Khalil T. (eds.) *Productivity Enhancement for Social Advance: The Role of Management of Technology – IAMOT 14th International Conference on Management of Technology Proceedings*, Vienna, Austria, 22-26 May.

Debackere, K. (1999) *Technologies to develop technology*. Antwerp, Belgium: MAKLU Publishers.

Eisenhardt, K.M. (1989) Building theories from case study research. *Academy of Management Review*, 14, 4, 532-550.

European Commission (2004a) *Nanotechnology, innovation for tomorrow's world*, April, available at http://www.cordis.lu/nanotechnology/src/pressroom.htm.

European Commission (2004b) *Vision 2020, Nanoelectronics at the center of change*, June, available at http://www.cordis.lu/nanotechnology/src/pressroom.htm.

European Commission (2004c) *Communication from the Commission: Towards a European strategy for nanotechnology*. Bruxells, Belgium, May.

Hellström, T. & Sjölander, S. (2005) *Entrepreneurial learning & academic spin-offs - Project report to Nordic Innovation Centre*. Göteborg: Norden Nordic Innovation Centre.

Hewson, C., Yule, P., Laurent, D. & Vogel, C. (2003) *Internet Research Methods.* London: Sage Publications.

Nanotech/IT. (2004) *Nanotech IT Newsletter*, 1, April.

Patton, M.Q. (1987) *How to use qualitative methods in evaluation.* Thousand Oaks, California, US: Sage Publications.

Shafer, S.M., Smith, H.J. & Linder, J.C. (2005) The power of business models, *Business Horizons*, 48, 199-207.

U.S. Government (2003) *21st Century Nanotechnology Research and Development Act*, December, available at http://www.nsf.gov/crssprgm/nano/activities/congress.doc.

U.S. Government (2006) *American Competitiveness Initiative. Leading the World in the Innovation*, Domestic Policy Council Office of Science and Technology Policy, February.

Yin, R. K. (2003) *Case Study Research: Design and Methods*. Thousand Oaks, California, US: Sage Publications.

The main websites consulted for the elaboration of the paper are[1]:

www.luxresearchinc.com, 2004.
www.nanobusiness.org, 2000.
www.nanoinvestors.com, 2004.
www.nanoinvestornews.com, 2004.
www.investinitaly.com, 2004.
www.nanotec.it, 2004.
www.nanovip.com, 2004.

[1]The list of websites includes only the ones cited in the text. The elaboration of the paper, in fact, required the consultation of a great number of websites since part of the empirical study (the extensive analysis) was mainly based on the web.

12

Environmental Approaches towards Industrial Company Management in the Czech Republic

Lilia Dvořáková and Tereza Kadlecová
University of West Bohemia in Pilsen
Czech Republic

1. Introduction

Consumption is a way of life typical for most of the world nowadays. Companies not only produce economic value but, through their production and consumption, they contribute substantially to environmental pollution and damage. Particular States and companies are pressurized to enhance their economic growth. However, steady economic growth brings with it a lot of negative effects - environmental damage being one of them. Nowadays, the issue of CO_2 reduction is at forefront of discussions in the European Commission and other contracting states of the Kyoto Protocol. Company activities need to be regulated by an environmental legislation on international as well as national level. On the other hand, company's environmental behaviour is, to a certain extent, a question of its´ self-determination and is dependent on managerial decisions. Companies can conform to the legislation only (reactive strategy) or can be voluntarily environmentally-proactive beyond the remit of legislation (proactive strategy). This chapter deals with the proactive approach towards the protection of the environment and how this is applied through so called voluntary environmental instruments by Czech industrial companies. Sections two and three introduce the proactive concept and the level of its adoption within the business environment. In section four, results of research into the `environmental approach to production and business activities in a company`, conducted by the authors of this chapter, will be presented.

2. Reactive vs. preventive strategies of environmental conservation

Many forms of preventive as well as reactive strategies for environmental conservation are being used in industrial engineering. As the name indicates the preventive strategies endeavour to prevent the origin of damage and seek for sources of pollution and waste. Preventive strategies have more potential and their realization should be supported. (Hyrslova, 2002) The reason why reactive strategies are not so effective or promising is to do with the fact they do not focus on the sources of environmental damage but only try to mitigate the negative consequences of production. The reactive approach is applied through implementation of so called "end-of-pipe" technologies that are as follows: for example, refuse compactors, collection containers and vehicles, waste heat recovery systems, air pollution filters, noise abatement investments and sewage treatment plants. As a result, the quantity of toxic agents in one environmental domain drops but rises in another domain.

Companies, as mentioned above, can be voluntarily environmentally-proactive beyond the remit of legislation that is becoming more and more stringent. Companies (no matter what size they are) have a range of voluntary environmental instruments at their disposal, that are of a preventive character and endeavour to find the sources of raw material and energy wastage and to do away with the causes of environmental pollution as a consequence of production. Table 1 provides an overview and comparison of chosen voluntary environmental instruments. Particular instruments differ in benefits for a company as well as expenses connected with their implementation.

Comparison criterion	Voluntary methods and instruments						
	EMAS	EMS/ ISO 14001	EMA	Ecodesign	LCA	Evaluation of Cleaner Production possibilities	Eco-labelling
Purpose	Regulative	Regulative Educative	Informative	Regulative	Informative	Informative	Regulative
Focus	Systems	Systems	Processes	Products	Products	Processes	Products
Normalization	Yes	Yes	No	No	Yes	No	Yes
Necessary external assistance	Yes	Yes	No	No	No	No	Yes
Preventive strategy	Yes	Yes	No	Yes	Yes	Yes	can be
Financial claim connected with an implementation	Yes	Yes	No	Yes-Conside-rable	Yes-Conside-rable	No	Yes
Labour input intensity	No	No	Yes	Yes-Conside-rable	Yes-Conside-rable	Yes	No
Economic benefits	Yes, partly	Yes, partly	informative benefits more likely	Yes	No	Yes – considerable	uncertain
Dedicated to	All company types	All company types	All company types	Manufac - turing companies	All company types	All company types	Companies with products/ services included in existing product categories
Logo/certificate	Yes	Yes	No	No	No	No	Yes

Table 1. Comparison of chosen voluntary environmental instruments

Voluntary environmental instruments are not only a significant tool for enhancing production efficiency and competitiveness, but also a forceful instrument of environment preservation. This double effect used to be called the win-win principle (environmental and economic). However, this win-win principle is not typical for reactive strategy within which end-of-pipe technologies are implemented. Investment for environmental preservation can be thus divided into two groups according to technology used, they are:

- Investment in integrated facilities (dedicated to pollution prevention),
- Investment in end-of-pipe facilities (dedicated to pollution removal).

Statistical data on `investment for environmental preservation`, related to business sector, has been available since 2003, whereas data for a public sector has been available only since 2006. Statistical data regarding environmental investment according to technology type has been monitored within the new ascertaining (ŽP 1-01) and the public sector has been interviewed since 2006. Some of the respondents do not structure their investments into the above categories and a small portion (to 3.5 %) of environmental investments stay therefore unclassified. In figure 1, a comparison of investment volumes into integrated and end-of-pipe technologies is displayed, for both business and public sector. (Kozouskova, 2008)

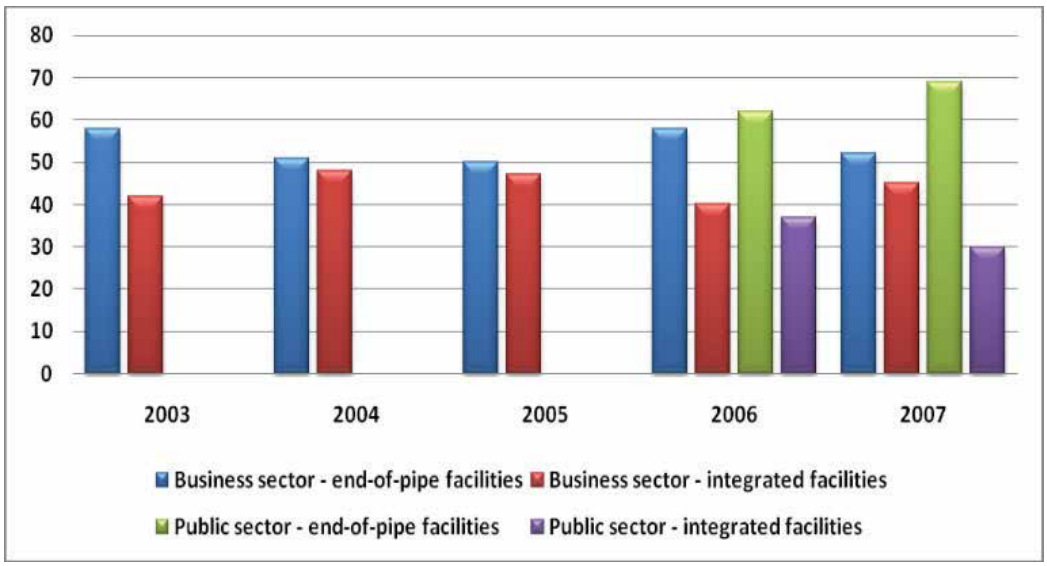


Fig. 1. Types of investment for environmental preservation, (Kozouskova, 2008)

According to the Czech Statistical Office investigations carried out in 2006 and 2007, companies implementing only the end-of-pipe technologies still prevail in the Czech Republic. (Kozouskova, 2008) In 2006, the total investment for environment conservation was about 13, 076 mil. CZK, whereas investment volume amounted to 57.8 % for end-of-pipe technologies and 40.9 % for preventive technologies. In 2008, the situation seemed to be slightly more favourable, the total investment for environment preservation rose to 14, 208 mil. CZK, the portion of end-of-pipe technologies was approximately 52.1 % and portion of preventive technologies rose to 44.6 %. (Vlckova, 2004) Conversion from corrective measures towards prevention is indisputably a positive trend, however, end-of-pipe technologies remain to be the only instrument for environmental preservation in many companies, despite the fact they are very expensive, their operation is very costly and effect for the environment is insufficient. Many companies are still afraid of the enhanced costs connected with the application of environmental instruments. Environmental conservation is perceived as a costly issue obstructing the economic performance that is therefore not worth adopting. On the top of that, according to the investigations carried out by the Ministry of Environment in the Czech Republic only 45 % of companies are able to enumerate the total costs for the system implementation and total revenues. The efficiency is

difficult to calculate. (Vlckova, 2004) In section 3, the most important proactive instruments are characterised briefly and the level of their adoption by Czech industrial companies is described as well.

3. Voluntary environmental instruments and their usage in the Czech industrial companies

3.1 Environmental management systems (EMS)

At present there are two standards for environmental management system implementation:

- Technical standard ISO 14 000; and
- Regulation (EC) No 761/2001 of the European Parliament and of the Council of 19 March 2001 allowing voluntary participation by organizations in a Community eco-management and audit scheme (EMAS).

The implementation of these EMS standards differ in the number of requirements specified by each standard. In ISO 14 000, some of the items required in EMAS are merely recommended, or sometimes not even specified. The main differences in the requirements of both above systems are shown in table 2. Because of the stricter requirements imposed by EMAS, only 28 Czech companies are certified according to EMAS; whereas about 3000 Czech companies have obtained certification according to ISO 14 000. In figures 2 and 3, a trend in number of Czech companies having implemented EMAS and ISO 14001 is displayed.

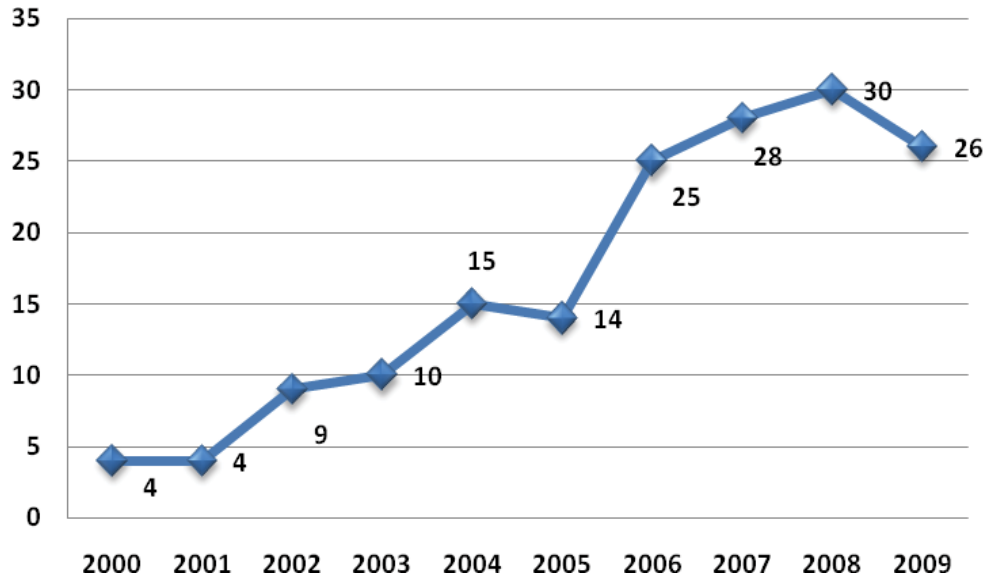


Fig. 2. Companies with EMAS implemented in years 2000 – 2009, (Cenia, 2009)

3.2 Environmental management accounting

UN Division for Sustainable Development, Expert Working Group on "Improving the role of Government in the Promotion of Environmental Managerial Accounting" defines the environmental management accounting as follows: Environmental management accounting (EMA in the following) is integral instrument to company management enabling

identification of environmental costs and revenues. Within EMA, financial flows as well as physical and energy flows are observed; EMA thus consists of two subsystems: (Jasch, 2002)

1. Monetary EMA – MEMA; and
2. Physical EMA – PEMA.

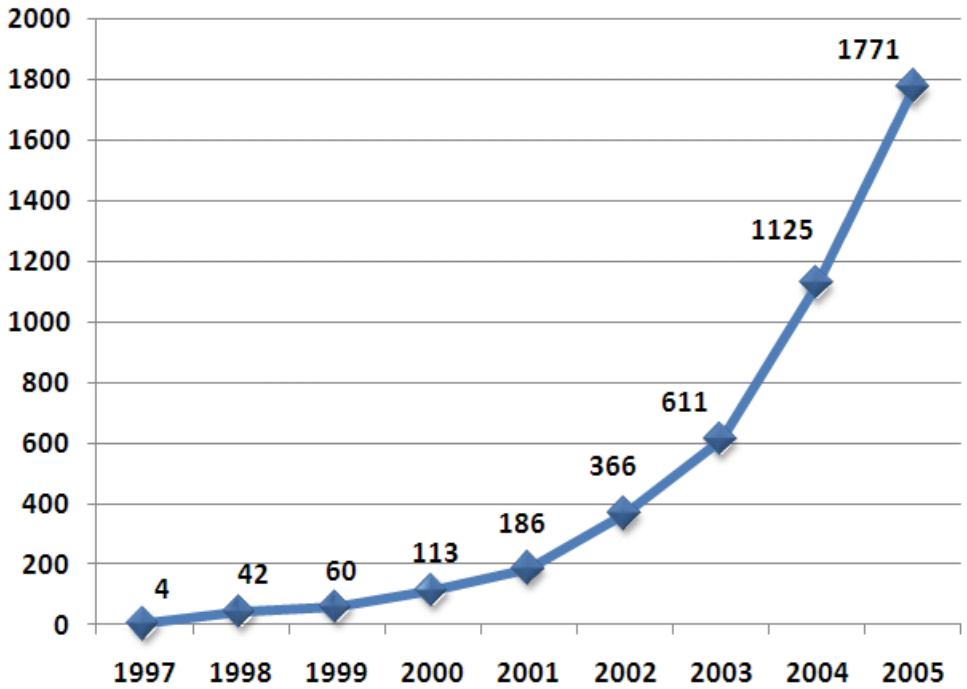


Fig. 3. Companies with ISO 14001 implemented in years 1997 – 2005, (Klasterka, 2007)

Among the main tasks of EMA are cost identification, data collection, and the preparation of estimations, analyses and reports. Collectively these activities contribute: (Jasch, 2002)

- information about material and energy flows;
- information about environmental costs; and
- other information expressed in values that are the starting point for decision making in a company.

Within EMA the information from both subsystems is interconnected and the value aspect of material and energy flows is expressed. The information that is monitored relates not only to a company as a whole, it can even relate to particular business processes, sections and plants. Among the aims of EMA are: (Jasch, 2002)

- enhancing efficiency of material and energy utilization;
- mitigation of environmental impact of company activities, products and services;
- mitigation of environmental risks; and
- trading income improvement.

Many companies are not familiar with the concept of environmental accounting, this refers especially to small and medium-sized enterprises (SME) which show little interest in this issue. They work to keep within regulatory limits and to minimize their environmental costs. This problem is aggravated by the absence of a simple methodology to monitor the environmental costs and revenues in SMEs.

----------	ISO 14001	EMAS
Force	worldwide	EU members
Acceptance	all kinds of companies (e.g. industry, services, public service)	all organizations that have impact on the environment
Implementation	economically separate parts of a company or a company as a whole	in the whole company area
Introductory environmental review	not requested but recommended	compulsory
Public documents	only environmental policy	environmental policy and the declaration about the state of the environment
Environmental declaration	none	is requested
End of the process	certification	verification of declaration about the state of the environment
End of the process ensured by	auditor of a certification company	accredited environmental verifier
Frequency of audit	not provided	three-year most
Logo usage	none (except the certification authority logo after an agreement)	usage of EMAS logo
Registration	in terms of issued certifications by particular certification organizations	corresponding subjects of particular member states

Table 2. Comparison of EMS according to ISO 14001 and EMAS, (Cenia, 2006)

The situation is much better in larger companies which often have a proactive attitude to protection of the environment. They often compile their own environmental accounting system to manage their high environmental investments and significant operating costs. (Ruzicka, 2002)

3.3 Cleaner production

The concept of cleaner production is being connected with the integral preventive strategy which is applied especially to production. The aim of this strategy is to do away with the causes of environmental pollution as a consequence of production. All the company processes are observed as a whole in terms of their impact on all domains of the environment. It is therefore not possible to transfer the negative impact from one domain of the environment into another one as it is with the end-of-pipe technologies. In order for the sources of undesirable waste to be identified, material and energy flows are monitored within cleaner production.

Afterwards the possibilities of elimination of these sources are explored, namely: (Remtova, 2003 - a)

1. Ease of technical feasibility;

2. Final economic efficiency;
3. Environmental efficiency.

The implementation of cleaner technology in a company is not a one-shot action but a long-term process. Effectiveness is the reason why a company should deal with cleaner production. In other words, reducing raw material and energy consumption reduces the negative impact on the environment. Waste increase is thus being prevented at the source and this leads to a significant economic effect at the same time. The negative impact on the environment can be reduced by engineering as well as non-technical (organizational) methods, which are very effective and are connected with none or very low costs in many cases. In the Czech Republic, cleaner production activities were introduced in 1992 within a Czech-Norwegian project dedicated to establish the fundamental facilities for widening the concept of cleaner production. The Czech Republic adopted the cleaner production strategy officially in 1999 when the International Cleaner Production Declaration was signed. Later on in 2000 the Cleaner Production National Programme (NPCP) was adopted. (Cenia, 2005) An essential amount of the projects was assured by the Czech Centre of Cleaner Production (CPC), and by the Cleaner Production Centre in Brno city. However, the application level of cleaner production is in practice rather low, as can be seen in the figure 4.

3.4 Eco-design

Eco-design is one of the preventive oriented voluntary regulative instruments of environmental policy. While cleaner production focuses on a company as a whole, the concept of eco-design concentrates on product development and design. Eco-design incorporates requirements of environmental protection into product design and development.

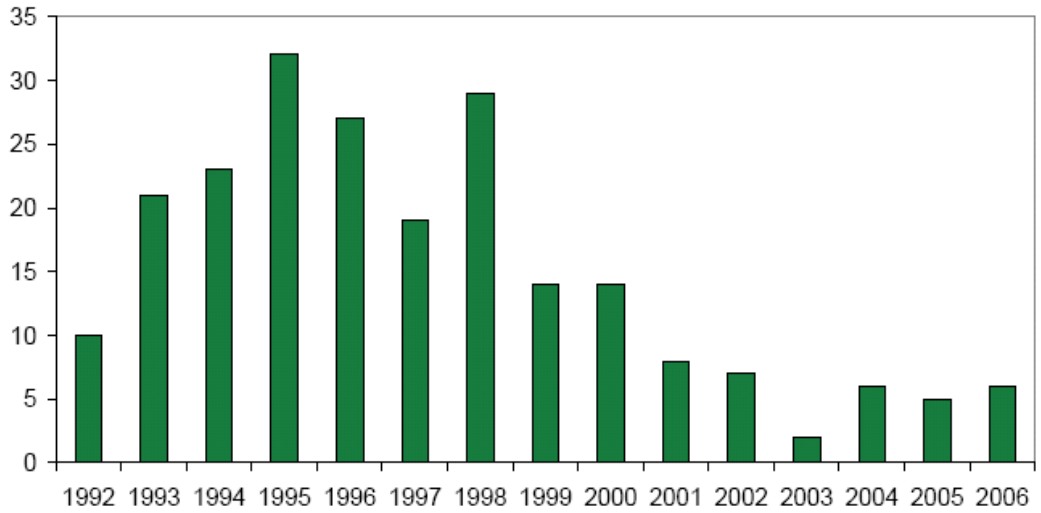


Fig. 4. Number of cleaner production projects in years 1992 – 2006, (Cenia, 2008)

So far there is no unified definition of eco-design. In general, eco-design can be defined as a systematic process of product design and development which puts emphasis not only on classical features such as functionality, economics, safety, ergonomics, technical feasibility,

aesthetics, but it also emphasize the minimum negative impact of a product on the environment during its whole life cycle. (Remtova, 2003 -b)
Eco-design activities are mainly concerned with: (CIR, 2004)

- substitution of dangerous materials with less dangerous ones,
- implementing measures leading to waste minimization,
- reduction in raw material consumption,
- packing and transportation optimization.

Eco-design is not an unknown concept to Czech engineering designers and some manufacturing corporations. Product innovation volume in the Czech Republic, however, is unsatisfactory. On the basis of the Innovation and Development Centre investigation from 2004 it is obvious that the total investment in development and innovative technologies in companies was only 48 billion Czech crowns, which is less than 2 % of the sales in all innovating companies. According to these investments 45 % was spent on new technologies and equipment, but only 2 % was invested in design projects. (CIR, 2004)

3.5 Life Cycle Assessment (LCA)

LCA is one of the most important information instruments of an environmental oriented production policy. LCA evolved from the Resource and Environmental Profile Analysis (REPA) in the USA during the late 60´s and 70´s of the 20th century. This method focused on product evaluation in terms of energy and raw material consumption. (Remtova, 2006) In order for the negative impacts of a particular system on the environment to be determined, inputs (material and energy flows) for production and outputs (production and services, waste) to the environment are compared within LCA.
LCA provides information about the impacts of a product in terms of its whole life. This method considers emissions in all domains of the environment during the production, utilization and product disposal phase. Effects of other processes related to acquiring raw material, material and energy production are involved as well. The structure and procedure of LCA is strictly determined by the International Standard ISO 14040 that defines the LCA as `gathering and evaluation of inputs, outputs and potential impacts on the environment during the whole life cycle of a product system`. Commercially available databases of processes as well as material and energy flows are being used for effective LCA studies processing.

3.6 Eco-labelling (environmental declaration type I)

Eco-labelling is a system that certifies that specific products and services have less negative impacts on the environment than their competitors and are, therefore, friendlier to the environment. This system is directed by an independent third party. Nowadays there are more than 30 eco-labelling systems and their number is increasing. National or supranational labels are assigned within those systems. Eco-labelling is regulated by ISO 14024. The products and services that apply for the certification have to meet many requirements concerning quality of the product (service); particular production phases; use of raw material and technologies, and final disposal. In contrast to the majority of preventive strategies focusing on systematic examination of manufacturing processes, eco-labelling makes use of market mechanisms outside the company which are based on supply and demand. In the Czech Republic, the eco-labelling system was established due to an initiative of the minister of Environment and the minister of Economy. In 1994, the National Program of Environmental friendly product was declared in the decree of the government.

The trade mark is a property of the Czech Ecological Institute (ČEU). (Ekoznaceni, 2003) In the Czech Republic, the eco-labelling system is executed as the National Programme of Environmentally Friendly Products Labelling, respectively on the basis of the Government Decree Nr. 159 from April 7, 1993.

However, the environmentally friendly products market has so far not evolved very much in the Czech Republic, economic return is therefore relatively low. The eco-labelling eco-efficiency is rather poor, despite the fact that the Czech eco-labelling programme is so far much cheaper than foreign programs, as a registration fee of only 735 EUR needs to be paid. (Vlckova, 2004) The figure 5 shows the number of eco-labelling licences awarded between years 1994 to 2004.

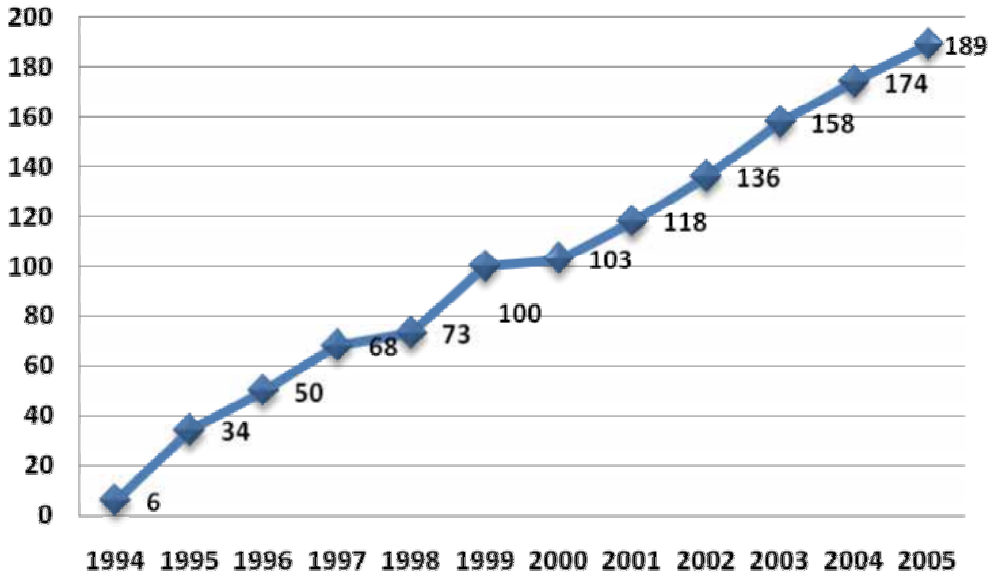


Fig. 5. Number of Czech eco-labels awarded between years 1994 and 2005

4. Research regarding environmental approach to production and business activities in a company

4.1 Characterization of the research and the surveyed sample

At the beginning of 2009 (from January to April), the authors conducted, along with the Institute of industrial management, Ltd., research on an "Environmental approach to production and business activities in a company". 150 mechanical engineering companies located in the western part of the Czech Republic, randomly selected, were sent questionnaires, from which 37 companies (24.7 %) returned a completed questionnaire. In the questionnaire, questions were asked regarding companies' attitude to environmental activities, purchasing of green products, certification according to ISO 14001 and EMAS or application of environmental instruments.

The interviewed companies were classified into three groups (small, middle and large) according to criteria of the Industry Association of the Czech Republic. 21 companies (57%) represented large companies (501 and more employees and turnover over 101 millions

CZK), 10 companies (27 %) represented middle companies (101-500 employees and turnover above 101 millions CZK) and 6 companies (16 %) represented small companies (1 – 100 employees and turnover to 30 millions CZK) (see figure 6).
58 % of interviewed companies were all-Czech, whereas 29 % of companies stated that the bulk of their capital was foreign (see figure. 7). On the basis of the existing situation, the voluntary environmental activities were anticipated to be applied especially in the larger companies with the bulk of the foreign capital share. This presumption, however, has not been confirmed by the research conducted. No clear connection between foreign capital share in a company and the company attitude towards the environmental activities has emerged from the research. 90 % of companies with bulk of the foreign capital stock acknowledged a passive approach towards the environment conservation within their company activities.

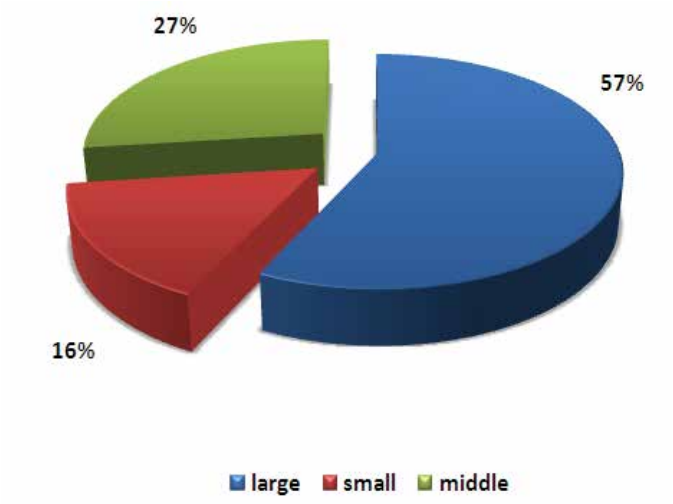


Fig. 6. Size structure of the companies

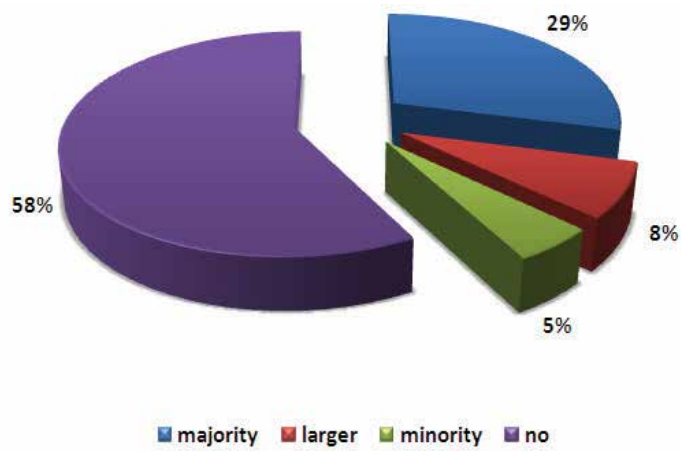


Fig. 7. Share of foreign capital stock in the companies

4.2 Research outcomes

As the research showed, only 19.4 % (big companies only) of all interviewed companies are voluntarily environmentally-proactive beyond the remit of legislation. 81.6 % of all the companies interviewed conform to the environmental legislation only (68.2 % of the large companies, 100 % of the small and middle companies). From the table 3 it is obvious that end-of-pipe technologies prevail in Czech industrial companies. This relates especially to small and medium enterprises (SME). Larger companies implement some of the environmental voluntary instruments like cleaner production (24.2 % enterprises), environmental management accounting (6.1 %), eco-design (6.1 %) and other.

Cleaner production seems to be popular, to some extent, to small and middle enterprises as well. Eco-labelling and LCA is not being used by the interviewed companies at all. According to the research outcomes, 42.1 % of companies (large companies mostly) have implemented monitoring system of costs and revenues relating to environmental issue. The situation is obvious from the figure 8. Only 6.1 % of companies interviewed make use of environmental management accounting (EMA). It can be concluded that companies do have their own costs and revenues monitoring system, however, they do not use the information instrument EMA very often. Lack of interest in environmental friendly products indicates the poor relation of Czech enterprises to eco-labelling as well. 34.2 % of interviewed companies do not buy green products and services at all, 28.9 % buy these products rarely and 36.8 % buy these products from time to time. No enterprise buys the green product regularly. This is a closed cycle: no demand, no supply. Poor green products demand is probably the reason why companies do not produce and offer environmental friendly products and services. According to this research 15.8 % of companies deem it necessary that their business partners have implemented EMS, for 28.9 % of companies it is less important if their business partners implement EMS and 10.5 % of companies do not take into account EMS implementation when seeking business partners.

Environmental methods and instruments	Company size			
	Small	Middle	Large	Total
End-of-pipe technologies	60%	81.8%	57.6%	63.3%
Environmental management accounting	20%	0%	6.1%	6.1%
Cleaner production	20%	9.1%	24.2%	20.4%
Eco-design	0%	0%	6.1%	4.1%
Eco-labelling	0%	0%	0%	0%
LCA	0%	0%	0%	0%
Other	0%	9.1%	6.1%	6.1%
Total	100%	100%	100%	100%

Table 3. Usage of chosen methods and instruments for environmental conservation

Companies´ attitude to environmental activities is interesting as well. Environmental activities are perceived by one half of the small companies interviewed as costly and providing hardly any benefits. 60 % of middle companies believe, on the contrary,

environmental activities to be prestigious and beneficial to business. This view is shared by almost 91 % of all large enterprises. Figure 9 demonstrates this situation. Regarding environmental management systems, no small company interviewed is certified according to EMAS or ISO 14001. 60 % of middle companies do not acquire any EMS certification, 30 % are certified according to ISO 9001, 10 % according to ISO 14001. 66.7 % of large companies are certified according to ISO 14001, 8.3 % acquire other type of a certification (mostly ISO 9001), 16.7 % are not certified at all. However, many companies confess the certification means only a formality to them – "scrap of paper". The above situation is summed up in the table 4. The conducted research has acknowledged the conclusions of investigations undertaken by the Czech Statistical Office, Innovation and Development Centre and Cenia Agency (mentioned in sections 2 and 3) that the voluntary environmental instruments are not sufficiently used in the Czech companies.

Type of certification	Company size			
	Small	Middle	Large	Total
EMAS	0%	0%	8.3%	5%
ISO 14000	0%	10%	66.7%	43%
Neither	83%	60%	16.7%	38%
Other	17%	30%	8.3%	15%
Total	100%	100%	100%	100%

Table 4. EMS certification in companies according to their size

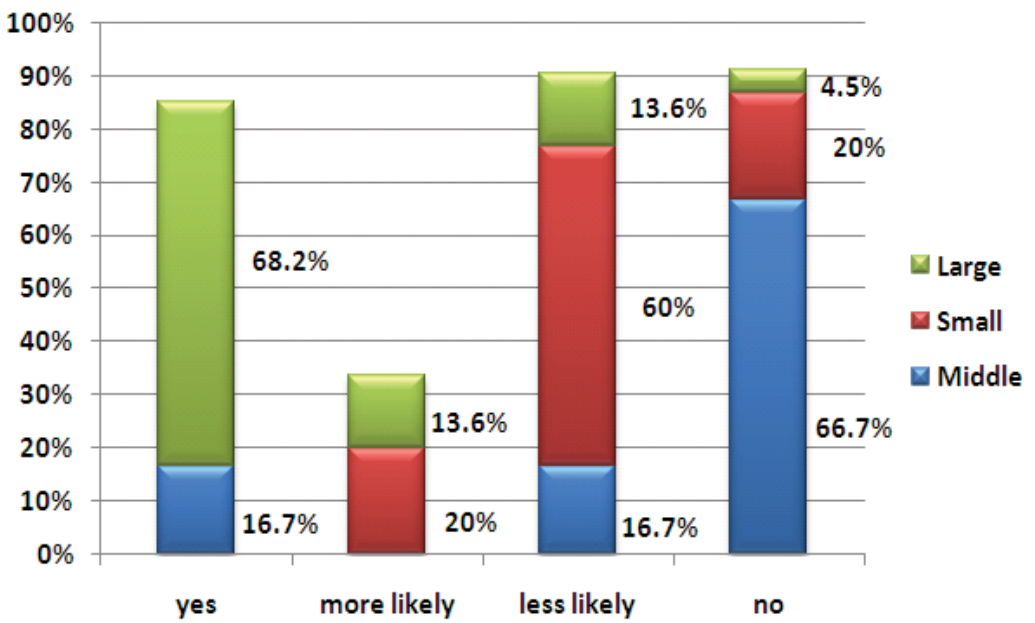


Fig. 8. Level of environmental costs and revenues monitoring according to company size

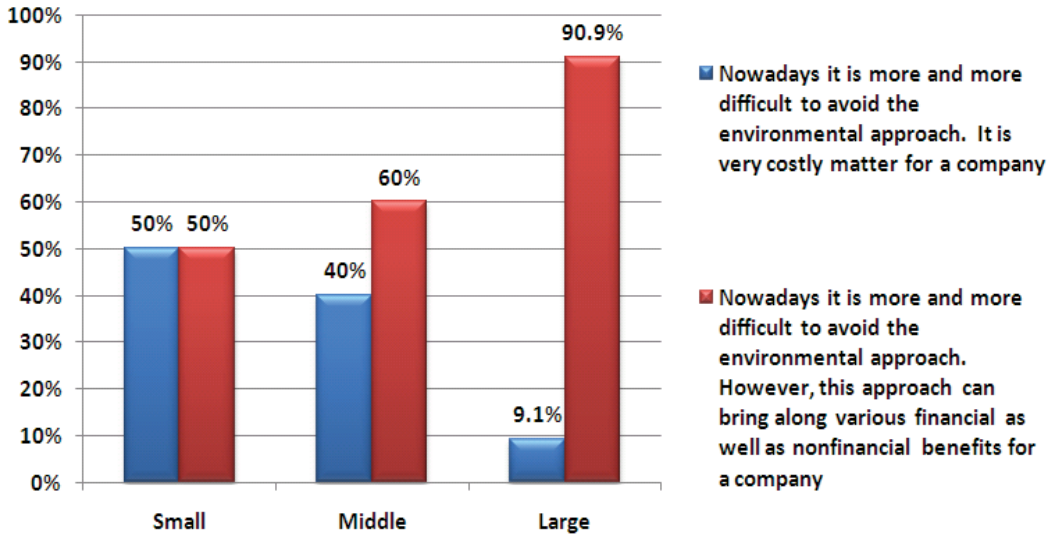


Fig. 9. Attitude to environmental activities in a company

4.3 Questionnaire

In this sub-section the questionnaire used during the research is presented. The questionnaire was aimed to be as simplest as possible for the respondents, the volume therefore did not exceed two A4 pages and all questions were constructed to be very easy to fill in by ticking off the correct answers. The following text represents an English translation of the original questionnaire that was presented in Czech.

Ladies and gentlemen
We would like to ask you to take part in this questionnaire inquiry that is conducted by the Department of Industrial Engineering and Management at the Faculty of Mechanical Engineering, at the University of Bohemia in Pilsen, and is devoted to research and development purposes only. To fill in the questionnaire, please tick one of the options for every question. It will take you less than ten minutes. Other comments and explanations are, of course, very welcome. Thank you very much for you participation and support in this questionnaire inquiry

1. Has your company implemented monitoring system of costs and revenues relating to environmental issue?

☐ YES
☐ MORE LIKELY
☐ LESS LIKELY
☐ NO

2. Is you company certified according to:

☐ EMAS
☐ ISO 14000
☐ Neither
☐ Other certification

3. What is your company's approach towards conservation of the environment?

☐ We conform to the environmental legislation only.
☐ We are voluntarily environmentally-proactive beyond the remit of legislation. How?

4. What methods and instruments for environmental conservation are used within you company?

☐ end-of-pipe technologies (facilities for waste and pollution treatment, e.g. refuse compactors, collection containers and vehicles, waste heat recovery systems, air pollution filters, noise abatement investments and sewage treatment plants)
☐ Environmental management accounting
☐ Cleaner production
☐ Eco-design
☐ Eco-labelling
☐ LCA (Life Cycle Assessment)
☐ Other

5. Does your company buy the `green` products and services?

☐ YES
☐ MORE LIKELY
☐ LESS LIKELY
☐ NO

6. During the supplier selection procedure, is the environmental approach of the potential supplier essential for you decision? (E.g. ISO 14001, EMAS certification)

☐ YES
☐ MORE LIKELY
☐ LESS LIKELY
☐ NO

7. What statement does better express your company's attitude towards environmental issues?

☐ Nowadays, it is more and more difficult to avoid the environmental approach that is very costly issue for the company.
☐ Nowadays, it is more and more difficult to avoid the environmental approach; however this approach can bring along various financial and non-financial benefits to a company.

8. What is the portion of foreign capital share on the total capital of your company?

☐ Majority
☐ Larger
☐ Minority
☐ None

5. Conclusion

In the current business environment, companies need to weather many changes in terms of technical, economic, legislative, environmental and social conditions. If a company endeavours to hold out in the competition and to stay in the market, it has to respect its environmental and social responsibility for its surroundings. Companies have a range of environmental instruments that are not requested by law but are beneficial in many aspects

to company and society as well. So far, the approach of Czech industrial companies has been rather reluctant towards the environmental aspects of business that are perceived as an obstruction to their economic development. Voluntary environmental instruments are used especially by larger companies that hold the ISO 14001 certification. Small companies, on the contrary, are not often even aware of the existence of these instruments and cannot thus know how to make use of them. Among the main barriers to application of the voluntary environmental instruments are, for example, lack of financial funds, lack of qualified workforce, and lack of time. However, it can be stated that Czech industrial companies are awaking to their environmental responsibility and the benefits it can bring, they are slowly moving from taking corrective measures towards prevention applied through voluntary environmental instruments.

6. References

Cenia. (2005) *Programme for Environmental technologies promotion in the Czech Republic. Program podpory environmentálních technologií v České republice.* Prague. [online, accessed on 2010-17-03]. Available from: <http://www.cenia.cz/web/www/web-pub2.nsf/$pid/CENMSFQP2T8Q/$FILE/ETAP%20CR.pdf>

Cenia. (2006) *Differences between ISO 14001 and EMAS. Rozdíly mezi ISO 14001 a EMAS.* [online, accessed on 2010-20-05]. Available from: <http://www.cenia.cz/web/www/web-pub2.nsf/$pid/CENMSFZS9TOS/$FILE/rozdily_ISO_EMAS.pdf>

Cenia. (2008) *Summary of cleaner production projects implemented between 1992 and 2006. Přehled projektů čistčí produkce v období 1992 - 2006.* [online, accessed on 2010-11-04]. Available from: http://www.cenia.cz/web/www/web-pub2.nsf/$pid/ MZPMSFHODF0K/$FILE/Projekty%20CP.pdf

Cenia. (2009). *EMAS Statistics for 2009. Statistika EMAS za rok 2009.* [online, accessed on 2008-03-04]. Available from: <http://www.cenia.cz/web/www/web-pub2.nsf/$pid/CENMSFZSAHSJ/$FILE/2009.pdf>

CIR. (2004). *Situation Analysis on Ecodesign in the Czech Republic*, *Situační analýza o ekodesignu v České republice,* Innovation and Development Centre, Centrum inovací a rozvoje. [online, accessed on 2009-01-20], Available from: <http://www.cir.cz/ekodesign/482664/1833650>.

Ekoznačení. (2003). *Eko-labelling Publication. Publikace*. MŽP, ISBN 80-7212-223-1, Prague [online, accessed on 2008-06-13] Available from: <http://www.env.cz/osv/edice.nsf/AFE8148C8858BD4BC1256FF9003E2CD9/$file/EŠV.pdf>

Hyršlová, J. (2002). *Guideline for Implementation of Environmental Management Accounting. Metodický pokyn pro zavedení environmentálního manažerského účetnictví* Ministerstvo životního prostředí MŽP, Prague. [online, accessed on 2007-12-12], Available from: < http://www.oldmzp.cz.>

Jasch, Ch. (2002). *EnvironmentalManagement AccountingProcedures and Principles Způsob určování hodnot pro environmentální manažerské účetnictví*. Institute for Environmental Management and Economics. IÖW, Vienna. [online, accessed on 2007-12-12], Available from: < http://www.ioew.at.>

Klášterka, J., Růžička, P., Babička, L., Remtová, K. (2007). EMAS Systém environmentálního řízení a auditu - Přírůčka k Programu EMAS. Planeta, Vol. XIV., No 1/2007, p. 1-16, ISSN printed version 1801-6898 [online, accessed on 2010-03-04]. Available from: <http://www.cenia.cz/web/www/web-pub2.nsf/$pid/MZPMSFJ1TY3H/$FILE/planeta1_korektura2.pdf>

Kožoušková, E. (2008). *Investment into Environment Conservation. Investice na ochranu životního prostředí*. Czech Statistical Office, Prague. [online, accessed on 2010-17-05]. Available from: <http://www.czso.cz/csu/2008edicniplan.nsf/t/3D004C9D02/$File/200908c02.pdf>

Remtová, K. (2003) - a. *Cleaner Production. Čistší produkce*. MŽP, ISBN 80-7212-260-6, Prague [online, accessed on 2008-11-04]. Available from: <http://www.env.cz/osv/edice.nsf/e26dd68a7c931e61c1256fbe0033a4ee/820af3233682e83ec1256fc0004eaf10?OpenDocument>

Remtová, K. (2003) - b. *Eco-design. Ekodesign*. MŽP, ISBN 80-7212-230-4, Prague [online, accessed on 2008-06-08]. Available from: http://www.env.cz/osv/edice.nsf/da28f37425da72f7c12569e600723950/7907a38f19e1d57ec1256fc0004fe74d?OpenDocument>

Remtová, K. (2006). *Business Strategy for environment conservation, voluntary instruments. Strategie podniku v péči o životní prostředí, dobrovolné nástroje*. Oeconomica, ISBN 80-245-1086-3, Prague

Růžička, P. (2002). *Relationship between EMS/EMAS and environmental management accounting. Vztah EMS/EMAS a environmentálního manažerského účetnictví*. Informace o projektu Chemas. MŽP, Prague [online, accessed on 2008-03-04]. Available from: <http://www.env.cz/www/zamest.nsf/defc72941c223d62c12564b30064fdcc/a0f76376b58e1af1c1256d60003d56aa?OpenDocument>

Vlčková, J. (2004). *Eco-profit. Ekoprofit*. [online, accessed on 2008-11-04]. Available from: <http://www.ireas.cz/download/projekty/www_dns/priloha14.pdf>

13

Drilling Fluid Technology: Performances and Environmental Considerations

Mohamed Khodja[1], Malika Khodja-Saber[2], Jean Paul Canselier[3], Nathalie Cohaut[4] and Faïza Bergaya[4]
[1]*Sonatrach/Division Technologies et Innovation, Avenue du 1er Novembre, Boumerdès,*
[2]*Sonatrach/Division Laboratoires, Avenue du 1er Novembre, Boumerdès, 35000*
[3]*Université de Toulouse, INPT, UPS. Laboratoire de Génie Chimique UMR 5503 CNRS,*
4 Allée Emile Monso, BP84234, F31432 Toulouse Cedex 4
[4]*Centre de Recherche sur la Matière Divisée (CRMD) UMR 6619 CNRS,*
1b, Rue de la Férollerie 45071 Orléans Cedex 02
[1,2]*Algeria*
[3,4]*France*

1. Introduction

Petroleum drilling is the primordial step in the success of oilfield exploration. This success is based, on the one hand, on the important details derived from geological drilled formations and, on the other hand, on the good drill-in reservoir conditions. Thus, the paramount drilling objectives are to reach the target safely in the shortest possible time and at the lowest possible cost, with required additional sampling and evaluation constraints dictated by the particular application. Drilling the wellbore is the first and the most expensive step in the oil and gas industry. Expenditures for drilling represent 25% of the total oilfield exploitation cost and are concentrated mostly in exploration and development of well drilling. In the 90s, drilling operations represented about $10.9 billions, compared with $45.2 billions (API, 1991), the total cost of US petroleum industry exploration and production.

Drilling fluids, which represent till one fifth (15 to 18%) of the total cost of well petroleum drilling, must generally comply with three important requirements: they should be, i) easy to use, ii) not too expensive and iii) environmentally friendly. The complex drilling fluids play several functions simultaneously. They are intended to clean the well, hold the cuttings in suspension, prevent caving, ensure the tightness of the well wall, flood diesel oil or water and form an impermeable cake near the wellbore area. Moreover, they also have to cool and lubricate the tool, transfer the hydraulic power and carry information about the nature of the drilled formation by raising the cuttings from the bottom to the surface. Figure 1 shows a simple diagram of a rotary rig.

Drilling fluids went through major technological evolution, since the first operations performed in the US, using a simple mixture of water and clays, to complex mixtures of various specific organic and inorganic products used nowadays. These products improve fluid rheological properties and filtration capability, allowing to penetrate heterogeneous geological formations under the best conditions.

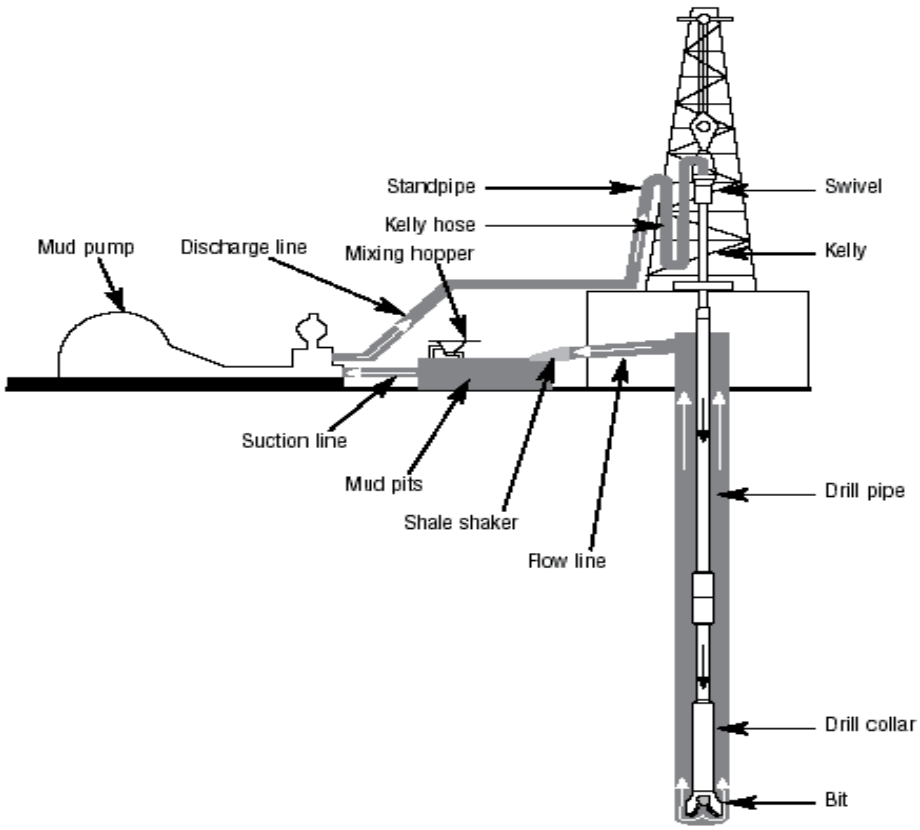


Fig. 1. A simple diagram of a rotary drill rig

In fact, borehole stability remains the main problem during drilling and the selection of drilling fluid type and composition was at the origin of successful drilling. Numerous studies have analyzed shale problems and several methods have been proposed to improve fluid performances for clay swelling inhibition (see §3 and 4) and to evaluate the scattered results already published in the literature. The majority of procedures recommend to compare initial and final sizes (or weights) of cuttings for inhibition estimation after fluid contact.
Then the question is to know which main factor (clay type, clay content or cuttings size) is affecting these disparate results.
In drilling fluid technologies, two main tendencies are currently developed in parallel: i) the search for new additives increasing the performances of water-based muds (WBM) and ii) the development and introduction of new compounds into oil-based muds (OBM). Some pendent questions will be discussed in this chapter, as well as filtration, formation damage and environmental considerations. Finally, some new solutions will be proposed by the authors.

2. History of drilling fluid technology

2.1 Drilling fluid composition

The complexity of the problems met in petroleum drilling has led to emerging techniques for the formulation of appropriate fluids. Generally, drilling muds may be classified in the following three families:

1. The WBM family, in which fresh-, salt-, or sea-water is the continuous phase, is the most used (90-95%). The WBM are mainly composed of aqueous solutions of polymers and clays in water or brines, with different types of additives incorporated to the aqueous solution.
2. The OBM family is less used (5-10%). These drilling fluids have been developed for situations where WBM were found inadequate (Chilingarian and Vorabutr, 1981). The OBM are oil- (usually, gas oil-) based muds. Generally, they are invert emulsions of brine into an oil major, continuous phase stabilized by surfactants. Also, other additives are often added to the organic phase, such as organophilic modifiers of the clay surface. However, although OBM often give better performances, they have major drawbacks such as to be generally more expensive and less ecologically friendly than WBM. Consequently, although OBM give greater shale stability than WBM (Bol et al., 1992), these latter systems have also been developed by many researchers in order to respond to environmental regulations (Simpson et al., 1994; Friedheim et al., 1999; Young and Maas, 2001; Patel et al., 2001; Schlemmer et al., 2002).
3. The third family of drilling fluids comprises gas, aerated muds (classical muds with nitrogen) or aqueous foams (Coussot et al., 2004). These drilling fluids are used when their pressure is lower than that exerted by the petroleum located in the pores of the rock formation. These fluids are called 'underbalanced fluids'. This underbalanced drilling technology is generally adopted for poorly consolidated and/or fractured formations.

Controlled drilling rate tests in various rocks have confirmed that air or gas is a faster drilling fluid than water or oil. Water should be the fastest drilling liquid, however, in this case, drilling tests show that the most commonly used additives have detrimental effects on the drilling rate. Choosing a mud system begins with the selection of a mud family, according to the nature of the rock formation, and should take into account environmental and economic constraints. The choice of the mud formulation will be the second step, where one has to decide on the range of desired properties, leading to use minimum amounts of additives. Figure 2 summarizes the drilling fluid types.

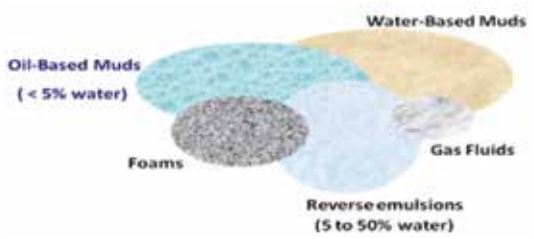


Fig. 2. Drilling Fluid Types

2.2 Biodegradability of drilling fluids

The biodegradability of petroleum products is dependent on the chemical structure of their various components. Compound resistance to biodegradation increases with increasing molecular weight. The oils used in OBM can be classified according to their aromatic hydrocarbon concentration, which contributes to fluid toxicity. However, the relations between hydrocarbon physico-chemical properties and biodegradability have been little studied. Several works (Zhanpeng et al., 2002), dealing with laboratory techniques of biodegradability determination and the influence of experimental conditions, showed the variation of the results according to the used method and considered conditions. In general,

the more soluble, lighter petroleum hydrocarbons are more biodegradable than the less soluble, heavier members of the group. Viscosity is also known to have an important impact on biodegradability. Highly viscous hydrocarbons are less biodegraded because of the inherent physical difficulty in establishing contact among contamination and microorganisms, nutrients, and electron acceptors compounds (Cole, 1994). The viscous diesel oil at high amount (>10%) shows low biodegradation rate (4%), but, in the presence of mixed culture (*Enterobacter sp., Citrobacter freundii, Erogenous Pseudomonas, Staphylococcus auricularis, Bacillus thuringiensis, Micrococcus varians,…)* it presents good biodegradation properties (Khodja, 2008). Moreover, the biodegradation behaviour of diesel oil does not obey that of individual compounds. With high amount of aromatics in diesel oil (33%), the difficulty was considerable to relate diesel oil biodegradability to its composition. Numerous works showed good correlation between biodegradability and some physical and chemical parameters. Haus et al. (2003) demonstrated that biodegradability decreased with increasing amounts of aromatic and/or polar compounds. He showed that kinematic viscosity is the significant factor in biodegradability variation with chemical composition and oil physical and chemical properties. Zhanpeng et al. (2002) based their method to calculate biodegradability on three parameters: BOD_5/COD (biological oxygen demand after 5 days/ chemical oxygen demand) ratio, CO_2 production and microorganism activity by ATP (adenosine triphosphate). On the chemical structure scale, some works (Hongwei et al., 2004) showed that biodegradability was a function of total energy and molecular diameter.

2.3 Drilling fluid technology

Drilling fluid technology is in constant evolution due to i) rapidly expanding needs due to more severe conditions, such as high temperature and pressure, tight gas and shale-gas reservoirs…, ii) increasing technical demands, such as increased lubricity requirements in air drilling and iii) growing restrictions on oil-based systems, such as environmental remediation. To comply with the new government regulations restricting the use of some technologies or practices, drilling fluid manufacturers have responded by developing acceptable alternatives. However, these solutions usually have substantial added costs and limitations that are sometimes prohibitive. In summary, drilling fluid development needs to encompass the design of new environmentally acceptable WBM and oil-like systems that will provide alternatives to OBM.

Such new drilling fluids should provide superior filtration control to minimize fluid invasion damaging permeable zones. The properties of the resultant mud cakes should prevent sticking of the drill pipe against the borehole wall due to differential-pressure. Particularly, in horizontal or high-angle wells, these new fluids should also provide adequate hole cleaning capabilities. The study of cuttings transport flow, air foam behavior and fluid viscoelastic behavior will help understanding and improving this process.

In order to attain greater efficiencies and cost savings, the main point in a R&D program is the consideration of all the consequential aspects of drilling technology (Drilling and Excavation Technologies, 1994). Such additional R&D should focus on the 'Development of environmentally-benign drilling fluids', designing non-toxic drilling fluids and foams as alternatives to toxic OBM which are moreover difficult to remove from the drill hole.

2.4 Optimization of drilling fluid performances

Drilling optimization in oilfields is usually formulated by using mathematical models. In these models, some parameters appear to be fundamental.

Fluid density

Density is the first parameter to consider. For desired densities greater or lower than 1, WBM or OBM can be used, respectively. The latter are recommended especially for clay formations where this density should be sufficient for drilling. Generally, for both WBM and OBM, mud weight (density) can be increased by adding various solids or soluble materials. Other undesirable solids issued from geological drilled formations are not easily removed but will be reduced to finer particles, which could have some adverse effects on mud properties. The way to avoid such undesirable phenomena is to use high-speed shale shakers. In additional stages, to remove finer solids down to the 1 μm range, these devices are equipped with 50- to 100-mesh screens, using desanders, desilters, mud cleaners, and centrifuges. Undesirable solids that are less than 1 μm can only be removed chemically using medium- to high-molecular-weight flocculants. In addition, some recommendations specify the effects of size on rheology and fluid performances. Solids less than 1 μm have 12 times more effect on drilling rate than larger particles (Lummus and Azar, 1986). For these solids less than 1 μm, the shearing stress required to start the fluid motion will be greater than for larger particles.

Viscosity

The second parameter to consider is viscosity. It is a general term used to define the internal friction generated by a fluid when a force is applied to cause it to flow. This internal friction is a result of the attraction between the molecules of a liquid and is related to a shear stress. The greater is the resistance to the shear stress, the greater is the viscosity. In fact, standard viscosity measurements do not define flow behavior within shear rate ranges imposed at the bit, annulus, and pits. The viscosity at the bit affects penetration rate, which will be better when viscosity is lower. The viscosity in the annulus affects hole cleaning efficiency and the viscosity in the pits influences the effectiveness of solids separation techniques.
Numerous additives are added to the formulation in order to reach optimized specific purposes which are sometimes contradictory. For example, mud has to be viscous enough in order to be able to lift the cuttings to the surface, but at the same time, viscosity must not be too high in order to minimize friction pressure loss.

Fluid loss

The loss of drilling fluids is the last considered parameter. It is generally defined as the volume of the drilling mud that passes into the formation through the filter cake formed during drilling. It is often minimized or prevented by blending the mud with additives. A number of factors affect the fluid-loss properties of a drilling fluid, including time, temperature, cake compressibility; but also the nature, amount and size of solids present in the drilling fluid.
In high-pressure and high-temperature environments, optimization of the above mentioned three parameters is essential to lighten instability problems when drilling through shale sections. Under these conditions, selection of suitable mud parameters can benefit from analyses that consider significant thermal and chemo–mechanical processes involved in shale–drilling fluid interactions.
Nevertheless, some other factors are not taken into consideration in these mathematical models. For instance, it has been widely experienced that random factors related to soil layers, drill bits, and surface equipment, greatly affect drilling performance. Optimization involves the post-appraisal of offset well records to determine the cost effectiveness of

elected variables, which include mud and bit types, weight on bit, and rotary speed. Stochastic models are introduced to describe such random effects. This more practical model provided a better characterization for real oilfield situations as compared with other deterministic models, and has been demonstrated to be more efficient in solving real design problems.

For drilling fluid additive evaluation, five important parameters have been proposed:

1. Main function and chemical nature,
2. Compatibility/salt tolerance with other additives and temperature limitations,
3. Recommended treatment range and cost,
4. History/success of using,
5. Interferences, damage and risk such as geological interpretation effects, formation damage, health safety and environment (HSE) and waste treatment.

2.5 Drilling costs

Remediation costs attributable to drilling are not easily estimated. The difficulty of access, the type of pollutant present, and the nature and time of derived treatment will influence the total cost. In oilfield operations, drilling costs typically account for 50 to 80% of exploration finding costs, and about 30 to 80% of subsequent field development costs (Drilling and Excavation Technologies, 1994). Typical costs for shallow hydrocarbon wells (up to 1,250-ft depth) drilled in the United States are about $27/ft (Anderson et al., 1991).

The boreholes required for environmental remediation will be shallow, so it might be expected that their cost will range between $20 to $30/ft, similar to shallow petroleum wells. However, special circumstances may increase these costs substantially. If the drilled solids contain toxic or radioactive substances, the drilling cost may increase dramatically because of the need to collect, document, and dispose the cuttings and to decontaminate drilling equipment.

3. Shale problems during drilling

The stability of a drilling fluid is generally guaranteed by its homogeneity after a long aging period. For OBM systems, a phase separation and a viscosity decrease are direct indications of drilling fluid degradation. In WBM, phase separation is also an indicator of mud instability. Figure 3 summarizes the evolution of drilling fluid state.

Mud viscosity affects the dispersion and the swelling of shales and decreases the diffusion velocity in porous medium. Muds with high viscosity and a minimum filtrate volume are preferred for inhibition efficiency, according to classical filtration equations. (see §5.1). In terms of its composition and properties, the mud column (i.e., the vertical column of drilling mud in the borehole) is a dynamic system whose characteristics are frequently changing dramatically both in time and space. Mud composition changes as shales migrate into the column and are dispersed into the mud, or by chemical interaction between the mud and the formation.

3.1 Shale instability

Wellbore instability is the largest source of trouble, waste of time and over costs during drilling. This serious problem mainly occurs in shales (principally clays), which represent 75% of all formations drilled by the oil and gas industry. The remaining 25% are composed of other minerals such as sand, salt, etc. The physical properties and behavior of shale exposed to a drilling fluid depend on the type and amount of clay in the shale.

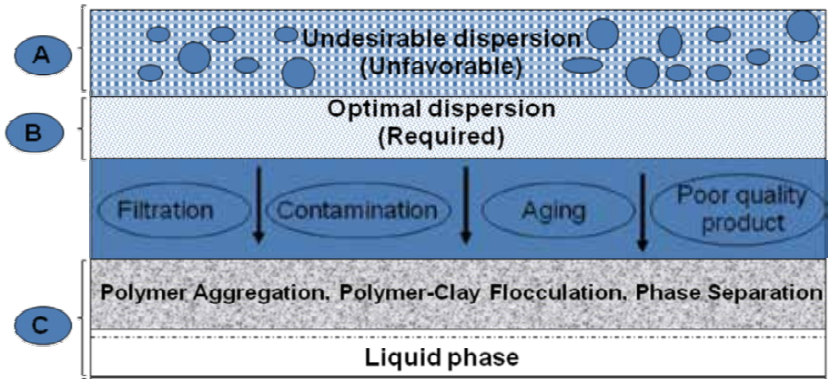


Fig. 3. Representation of drilling fluid destabilization (Khodja et al., 2010)
(Phase separation: low viscosity and high filtrate)
A: Undesirable dispersion with an inhomogeneous additive repartition. Solid-liquid, polymer-solution, dispersed phase-continuous phase are inhomogeneous and unstable.
B: Optimal dispersion with a uniform repartition of additives. The mud system is stable and exhibits good rheological and filtration characteristics.
C: The mud system is unstable for one of the following reasons: dramatic filtration conditions (pressure and temperature), use of incompatible additive (contaminant) or of poor quality products, or considerable aging. Solids, polymers, and salt in WBM, dispersed phase, emulsifiers or others additives in OBM are separated from continuous phase. The system presents a phase separation, involving a degradation of rheological parameters and a high filtrate volume.

Wellbore stability issues were not seriously addressed until the end of the 70s, when a famous published paper (Bradley, 1979) initiated great interest for this topic in the industry. Since that time, wells became more complex and drilling operations were routinely carried out in more difficult environments. In addition to a technical challenge, the occurrence of any wellbore instability-related problems will significantly add to the already high well costs. It is estimated that at least 10% of the well budget is used to perform unplanned operations resulting from wellbore instability. This cost may approach $1 billion/yr. worldwide.
Various aspects of wellbore instability have been presented recently. Shale–fluid interactions can be manipulated to enhance cuttings and wellbore stabilization as well as improving hole-making ability in shale formations (van 0ort, 2003). A membrane transport model was developed for calculating the diffusion potential and the reflection coefficient in shales under different conditions (Rosana et al., 2000). The ionic composition of the fluid saturating the shale appeared to control the magnitude of the membrane potential. This suggests that at least at early time, the type of cations in the drilling fluid is much less important than their concentration since this parameter controls the water activity. Thus, the stability of clay-rich shales is profoundly affected by their complex physical and chemical interactions with drilling fluids.
Most borehole stability and drilling fluid-related problems can be handled with present technology in well-defined environments if stringent quality control actions are maintained. Nevertheless, severe complex drilling situations still present serious challenges to economically viable drilling. Efforts and progress by several Companies have led to new

proprietary, or patented technologies usually available for license, applied in the field but rarely used in laboratories and scarcely published in the accessible literature.
When wellbore walls become unstable, the spilling of cuttings causes a disastrous change in the rheological properties of the mud (Beihoffer et al., 1988). Several studies on shale-fluid interactions confirm that various causes are at the origin of borehole instability: water adsorption, osmotic swelling and cation exchange.
Different approaches to WBM design have been suggested depending on given shale formations (Darley, 1969; Chenevert, 1970; Roehl and Hackett, 1982; Beihoffer et al., 1988; Zamora et al., 1990; Hale and Mody, 1992; Bol et al., 1992; Cook et al. 1993; Mody and Hale, 1993; Bailey et al., 1994; Simpson et al. 1994; Durand et al. 1995; Horsud et al., 1998; Pernot, 1999). Other recent studies focused on shale-fluid interactions (Lomba et al., 2000; Schlemmer et al., 2002; Van Oort, 2003). Consideration is given to maintain borehole stabilization in reactive shales by reducing hydration (swelling) and/or dispersion. This process is generally referred to as 'inhibition'. Clay wettability and inhibition properties were studied by analyzing the behaviour of water-clay-polymer-electrolyte systems. These properties are connected to the rheological and filtration characteristics for both mud and filtrate. Considering the replacement of OBM by WBM, Van Oort (2003) showed that additives, such as polymer and KCl, tend to reduce shale instability. Cuttings characterization is a key parameter to explain how salt, added to WBM, affects shale stabilization.
Although, as in most engineering disciplines, a wide gap appears between R&D studies and field applications, some important research areas could yield significant advances and benefits. In addition to new development, efforts should be made to transfer some of the old existing technologies that could immediately solve problems encountered in borehole stability or formation characterization and validation. The main critical parameters that should be determined are the constitution and strength of the formation, its discontinuities as well as abrasivity, permeability, pore pressure and stress state.
Studying clay stabilization problems, Kelley (1968) met contradictions, showing that some clay formations can be drilled easily meanwhile similar ones are dispersed. Several studies were conducted in order to understand these contradictory data and propose adequate mechanisms and solutions for drilling. Aadnoy (2003) intended to visualize that a good field modeling, based on the understanding of the underlying physics is the key for development of wellbore technologies and practices.

3.2 Clay swelling

Clay swelling is at the origin of well instability during drilling. Low and Anderson (1958) presented osmotic pressure equations for determination of the swelling properties, considering clays like semi-permeable membranes. Chenevert (1969) stating that the main reason of instability during drilling by WBM systems is the swelling of clays, adjusted the water activity of OBM systems, to prevent water adsorption on clays. Steiger (1993) studied clay hydration in a triaxial apparatus by measuring the swelling pressure of clays exposed to different drilling fluids with different water activities. He showed that the addition of potassium salts can reduce the water activity of clay and consequently the swelling pressure. In experiments conducted on site, he observed that the presence of KCl in the drilling fluid improves the stability of clay formations. Mody and Hale (1993) developed a model of stability supporting the interaction between drilling fluids and clay formation. This model identified the optimum drilling fluid parameters, such as density and salinity, for the elimination of instability problems during the use of WBM and OBM. They reported that the

chemical potential difference between water in the clays and in the drilling fluid is the most important parameter. Simpson et al. (1994), using an experimental approach, showed that OBM containing an emulsified water phase can prevent moisture and thus the weakening of the clay. According to these authors, the use of a hydrophilic organic compound, namely cyclic with multiple hydroxyl groups (methylglucoside) can also afford other characteristics similar to those of OBM, such as lubrication.

3.3 Laboratory methods for stability evaluation

Hale and Mody (1996) conducted experimental tests to study the direct impact of moisture on the mechanical properties of clay and tried to understand the mechanisms behind the instability of the wells. van Oort et al. (1996a) used the pressure transmission test (PT) (based on the work of Fritz and Marine, 1983) to measure the effectiveness of the clay membranes. They observed that, after increasing the upstream pressure, the outlet pressure increases due to a higher pressure in pore caused by the hydraulic flow. Horsud et al. (1998) have also studied the phenomenon of swelling pressure in clays and concluded that osmosis does not play a role, but that pressure (or suction) is the main parameter that controls the development of the swelling. Pernot (1999) quantified the effect of the swelling pressure of a variety of fluids in contact with several types of clays and concluded that the methylglucoside type 'Gumbo' stops clay swelling. The created barrier blocked the flow of ions and water in clays. Concentrated salt solutions show a low membrane reflection coefficient. Muniz et al. (2004) described the equipment used for the evaluation of clay-fluid interactions. The idea is to combine water and ionic gradients to estimate both the efficiency of the membrane reflectivity and the permeability coefficient and to integrate them into a program for stability evaluation. Zhang et al. (2006) developed the gravimetric swelling test (GST) and showed that water motion is not controlled only by osmosis (water activity) but is also influenced by capillary suction and ionic diffusion. The contact of fluid with clays changes their physico-chemical and mechanical properties.

Drilling fluid additives able to inhibit the swelling and dispersion of clays will be considered in §5.

4. Shale characterization and Inhibition

4.1 Inhibition diagnosis and shale characterization

The mechanism of inhibition is dependent on the choice of the polymer-salt system. It can be identified by the following features:

1. Increase of filtrate viscosity,
2. Reduction of clay permeability,
3. Balancing of the flow of mud filtrate in the clays with pore water by the effect of osmotic pressure ($a_{wdf} < a_{wsh}$), or,
4. Combination of the previous different factors.

Wellbore instability is due to the dispersion of the clay into ultra-fine colloidal particles and this has a direct impact on the drilling fluid properties. Clay characterization is the main parameter allowing understanding borehole stability.

Solid particles are divided into three groups according to size. Colloids from about 0.005 to 1 μm impart the viscous and filtration properties, silt and barite (sometimes called "inert solids") from 1 to 50 μm provide density, but are otherwise deleterious and sand from 50 to 420 μm, apart from bridging large opening in very porous formations, is objectionable

because of its abrasive property. Clay minerals are considered as particularly active colloids (Bergaya et al. 2006), partly because of their anisotropy due to shape (tiny platelets) and partly because of their molecular structure which presents high negative charges mainly on their basal surfaces, and possible positive charges on their edges. Interaction between these opposite charges strongly influences the viscosity of clay at low velocities, and is responsible for the formation of a reversible gel structure when the mud is at rest.

The main methods developed for shale characterization and fluid inhibition performances deal with composition, reactivity, mechanical and physico-chemical properties of shales (composed in majority of clay). A succinct list of usual methods is presented hereafter:

- **XRD,** X-ray diffraction analysis to determine qualitative mineral content,
- **CEC,** cation exchange capacity to evaluate reactivity of drilled cuttings. The methylene blue test (MBT) method was recommended by API 13I (2003),
- **GST,** a gravimetric swelling test, used to measure water and ion motion during shale/mud interaction (Zhang et al., 2004),
- **CST**, capillary suction time for the determination of filtration properties and salt optimization (Wilcox et al., 1987),
- **ROP**, rate of penetration measured with a penetrometer to estimate the degree and depth of softening (Reid et al., 1993) or with a Bulk Hardness Test designed to give an assessment of the hardness of shale following exposure to a test fluid (Patel et al., 2002),
- **DCM**, dielectric constant measurement to quantify swelling clay content and to determine specific area (Leung and Steig, 1992),
- **Triaxial test** for pore pressure measurements, carried out in downhole simulation cell (DSC) for compressive stress/strain behavior (Salisbury and Deem, 1990),
- **Oedometer test** for pore pressure modification and chemical potential influence (Bol et al., 1992),
- **SDT**, slake durability test, a standard method originally used in geotechnical studies when measuring the weathering and stability of rock slope: ASTM D 4644-97 (ASTM, 2000), reapproved 1992 (Likos et al., 2004),
- **Jar slake testing**, a qualitative method designed to evaluate shale relative durability in contact with a given fluid. Wood and Deo (1975), Lutton (1977) describe details of this method using six indices,
- **DSCA**, differential strain curve analysis for in situ measuring stress orientation and intensity (Fjaer, 1999),
- **Hot-rolling dispersion test** (shale disintegration resistance or cuttings dispersion test), the most widely used technique in optimizing drilling fluid. Appreciated for its simplicity, low cost and duration, it has been recommended by several laboratories and adopted by API 13B-1(2003).
- **Shale pellet inhibition** (pellet dispersion test): pellets and fluid are introduced into a steel bomb and processed as above (hot-rolling dispersion test). For comparison and reference, an OBM system is generally used (Mody and Hale, 1993).
- **Pressure transmission test**, used for confined or unconfined shale (van Oort, 1994). Muniz et al. (2004) described an apparatus designed to evaluate shale-drilling fluid interaction and estimate shale permeability, coefficient of reflectivity (membrane efficiency) as well as ionic diffusion coefficient,

- **Microbit drilling equipment**, requiring core sample availability and costly investment (Lamberti, 1999).

The comparison between all these techniques shows an important contribution of each of them. However, these methods are often criticized regarding feasibility, cost, precision and conditions used.

4.2 A new approach for inhibition evaluation

Swelling measurement is a key test when selecting and developing inhibitive WBM. However new methods are proposed, combining dispersion and pellet tests. The aim is to protect the initial quality of cuttings, to minimize grinding and to avoid moistening, while opting for a preliminary wash to eliminate the contamination of cuttings by the different additives (polymers, surfactants, etc.).

A new approach using a wet-cell X-ray diffraction method is proposed by the authors (Khodja, 2008). The advantage of this method is to evaluate clay swelling after fluid contact and to estimate differences in the rate of solution adsorption between various WBM systems. The principle is to combine in-situ X-ray diffraction in wet-cell with the evaluation of liquid adsorption. This latter method combines filtrate data (volume and rate) with rheological and inhibitive properties. The API fluid loss test (30 min, ΔP=100 psi through N°50 Whatman filter paper, ambient temperature) is the standard static filtration test used in the industry; however, because it uses very fine mesh paper as filter medium, all of the bridging particles are stopped at the surface of the paper and the spurt-loss phase is not simulated properly. A better static filtration test is the permeability plugging test (PPT), which uses a 1/4-inch-thick ceramic disk of known permeability (API 13B1, 2003). But in this test, mineralogy variation is not taken into account. In the new test, experiments were carried out by replacing Whatman 50 filter paper by the pellet in the API filtration cell (Khodja et al., 2008, 2010). The slurry was exposed to a 100 psi pressure for 30 min to obtain filtrate. The compaction force, linked to the deposit mode of the sediments, has a significant influence on the permeability.

With different systems (WBM with PHPA, glycol or silicate; OBM), our results show similar, rather high recovery values for large size (0.8 mm) but low recovery values for small size (0.100 to 0.315 mm) cuttings. When using different inhibitive polymers, almost no difference in recovered weight is noticed between cuttings samples from different geological formations and with different mineralogical compositions.

Our recommendation is then to use, in dispersion tests, preferably small size cuttings, which are in close contact with all additives used in drilling fluid systems. Moreover, when using small size cuttings, clays are fully exposed to the fluid and aggregation effect is eliminated (Khodja, 2008). Xanthan gum (or PAC) added as a viscosifier, acts synergistically with polyalkyleneglycols (PAG) and preserves cuttings integrity. To increase glycol efficiency, an inhibiting ion, preferably potassium**,** was used. For the silicate system, analyses show high adsorption of silicate ion on shale. The inhibition mechanism also depends on the type of polymer used, controlled by plugging of clay pores, thus reducing the dispersion (PAG), or by surface coating (film formation with PHPA or silicate).

Practically, drilling engineers need to optimize formulations in opposite ways depending on whether they deal with upper geological layers or reservoir formation. In the former case, minimum filtrate, optimal viscosity and high damage are required in fluid formulation selection. In the latter one, low damage is the principal selection parameter.

5. Drilling fluid additive evolution in WBM and OBM

Nearly a century after the birth of the drilling fluid industry, with hundreds of suppliers and thousands of manufactured products, water is still the main compound. Gas oil, initially a major technological breakthrough, has now been often replaced by synthetic low toxical oils (LTO) that lead to many problems and do not resolve critical drilling situations.

Crude starch and cellulose, the first used polymers, were constantly improved for thermal efficiency. Clays, historical bentonite additives, were first used in WBM. Small amounts of surfactants greatly modify drilling fluid performances. Nowadays, after treatment, bentonite is added to OBM under the form of organophilic clay (under the commercial name of Bentone). Clays, as additives, meet restrictions and regulations in accordance with environmental considerations, from the countries involved in oil drilling, or interested in drilling fluid research. This research, occurring at intensive laboratory scale and/or occasionally on site, is mainly based on experiments conducted in the case of drilling new discovered fields or of drilling in unknown geological sites. The less recommended use of OBM has renewed the interest in WBM, which also provide economic benefits. The major problem in the use of WBM is still linked to the instability of the wells, mainly due to the interaction of clays with the formation water, but several acceptable options are put forward. So, now, a large number of fluid systems are offered by specialized companies. In fact, a lot of these products are marketed, despite similar formulations appear under different tradenames.

Hereafter, some examples of WBM additives, which have improved the performance of drilling fluids, are presented.

Bentonite, a worldwide-used drilling fluid additive, is mainly a montmorillonite species. It is added to fresh water i) to increase hole cleaning properties, ii) to reduce water seepage or filtration into permeable formation, iii) to form a thin filter cake of low permeability, iv) to promote hole stability in poorly cemented formations, v) to viscosify the mud and finally vi) to avoid or to overcome loss of circulation. However, a low bentonite content is desired because a high clay content in drilling fluids shows several adverse effects, on the one hand, it greatly reduces the rate of penetration, and, on the other hand, it increases the chances of sticking due to differential pressure and it is the major cause of excessive torque and drag.

PHPA. Partially hydrolysed (30%) polyacrylamide is the most used additive in drilling for borehole stabilization in shale formations. PHPA-clay slurries tend to form a relatively thin filter cake at the borehole wall, characteristic often cited as an advantage (Darley and Gray, 1988).

KCl is a salt commonly used to inhibit the swelling of clays. According to van Oort (2003), the low efficiency of the membrane (1-2%) is probably due to the relative high mobility of KCl in the clays. In addition, conductivity and permeability are not altered and the osmotic pressure generated by KCl is moderate (typically less than 20 MPa). KCl-containing systems have good efficiency for the stabilization of clay cuttings in the presence of PHPA (Clark et al., 1976). Although Na^+ is not as good as K^+ ion, the use of NaCl has additional advantages. NaCl can reduce the invasion of filtrate into the clay. Indeed, close to saturation, NaCl leads to large viscosities and a water activity lower than those observed with concentrated solutions of KCl. In combination with silicates, polyols and methylglucoside, concentrated solutions of NaCl can improve the efficiency of the membrane (cake).

Amines and derived salts. Simple amines are used in several areas for specific applications. Quaternary ammonium salts prevent swelling and dispersion of clays by ion exchange.

Their disadvantages are their high cost, toxicity (Himel and Lee, 1951) and their incompatibility with anionic additives commonly used in fluids.

Anionic and non-ionic polymers. In order to stabilize clay particles and to prevent their swelling/dispersion behavior in the presence of water, some other ingredients are added. A wide variety of anionic (PAC: polyanionic cellulose), non-ionic (polyols, polyglycerols, glycosides, polyvinyl alcohol, hydroxyethylcellulose) or amphoteric polymers were tested. These polymers act by encapsulation, limiting water penetration in clays. However, they are generally less efficient for swelling than some cationic species (Stamatakis et al., 1995). PAC is used as a fluid loss reducer for fresh water and salt-water muds. Due to its anionic nature, adsorption and flocculation occur as a result of hydrogen bonding between solid surfaces and the hydroxyl groups on the polymer. The (poly-)glycerols and (poly-)glycols (Hale et al., 1989 and Perricone et al., 1998), usually simply referred to by 'glycerols' and 'glycols' have been widely used for drilling clays (Chenevert, 1989; Bland, 1991, 1992 and 1994; Downs et al. 1993; Bland et al., 1995). They prevent cuttings from dispersing into the medium (Bailey et al., 1994). Therefore, they increase drilling rates (Reid et al., 1993; Cliffe et al., 1995). Twynam et al. (1994) observed improvements with the use of a high concentration of glycol. Nair (2004) evaluated the performance of two commercial additives (Gilsonite® and Soltex®), respectively natural asphalt and high molecular weight modified hydrocarbon compound (sodium asphalt sulfonate), used as inhibitors in terms of high pressure and temperature (HP/HT), and got a small decrease in permeability without explaining the reasons for this reduction for both products.

Carbohydrates and derivatives. In response to environmental constraints, new families of compounds are proposed such as sugars and their derivatives (saccharides). Sugars increase the viscosity of the filtrate and reduce the flow of water in clays (van Oort, 1994). In addition, they provide a low water activity and generate an osmotic pressure favorable to clay dehydration. The problem with sugars is their susceptibility to biological attack, making them difficult to maintain unspoiled when stored on site. However, methylglucoside (MEG) and generally methylated saccharides are less susceptible to biological attack (Simpson et al., 1994). MEG is a derivative of glucose, supplied as liquid containing 70% solids. Made from corn starch, it is classified as "biodegradable". Saccharides are generally recommended for the stabilization of clays. Added salts to saccharide systems allowed effective dehydration of clays, reduction of "bit-balling" and increasing ROP. These MEG systems have a good filtrate and produce environmentally acceptable cuttings (Chenevert and Pernot, 1998). Soluble in water, MEG has many hydroxyl groups in a ring structure capable of reducing the water activity of the drilling fluid and may be a good additive to WBM.

Silicates and aluminum-based compounds have been introduced in the petroleum industry since the 90s (Ding et al. 1996; van Oort et al., 1996b; Ward and Williamson, 1996). They are strongly recommended for the stabilization of clays. Silicate-containing fluids show good shale swelling inhibition, low depletion rate and high ROP and, additionally, are environmentally friendly (Ward et al., 1997; van Oort et al., 1999; Tare and Mody, 2000). These soluble additives react rapidly with clays (Ca^{2+} and Mg^{2+}) to form insoluble precipitates by gelation which act as a barrier towards clay surface. The mechanism of gelation/precipitation can seal the micro-fractured clays (van Oort et al., 1996b). Compounds containing aluminum, 'Alplex™' (Clark and Saddok, 1993; Saddok et al., 1997), were also developed for this purpose.

The search for inhibitive WBM systems, which would perform like OBM, has been a continuous endeavor in the drilling industry. During the past several decades, many

approaches have been taken (Chenevert, 1970; Clark et al., 1976; Retz et al., 1991; Downs et al., 1993; Stamatakis et al., 1995), such as cationic polymer mud systems, glycol-based muds, polymer/salt systems, silicate muds, calcium ion-treated muds and other relatively high-concentration brine systems. However, all these approaches have not been completely successful in inhibiting the hydration of highly reactive systems and have various limitations. For example: i) cationic polymer systems are almost as inhibitive as OBM; however, the cost of running the system, the toxicity of cationic polymers, and their incompatibility with other anionic drilling fluid additives have resulted in limited success in the field; ii) highly-concentrated brine systems have limitations on mud formulations and properties; iii) while silicate muds have good inhibitive properties, they also pose problems related to human health and environmental issues, due to high pH values, logistic problems and mud formulation limitations; iv) some of the anionic and non-ionic polymer systems, e.g., biopolymers and PHPA-based systems, show limited thermal stability and mud formulation limitations.

6. Damage considerations. Petrophysical properties and filtration

6.1 Permeability/Porosity

In reservoirs, the solids found during drilling operation come from two principal sources: reservoir and drilling fluid additives. Generally, the formation damage directly apparent originates in the poor performances of un-weighted (without solid) fluids, giving low return permeability[1] (see §6.3). After fine tuning to achieve optimum particle size distribution (PSD) with a minimum solids, return permeability improved to high values. With bridging material, the invasion of the filtrate through the wellbore (spurt loss), which is a major damage mechanism, can be reduced to a minimum. Abrams (1977) recommended a minimum bridging particle concentration of 5% and a ratio of 1:3 between average pore size and medium particle size. Colloidal and hydrodynamic forces are responsible of fines liberation. Clays and fines are considered as producing one of the major damage of the formation. This damage is located near the wellbore area within a three to four feet radius. Dispersion and fines liberation in the majority of soils is promoted by high pH, high Na^+ saturation and low ionic force (Roy et Dzombak, 1996; Swartz et Gschwend, 1998 and 1999; Saiers et Hornberger, 1999; Grolimund et al., 2001). Kaolinite, in majority, and some illite exist in the formation rocks as pore filling materials. Kaolinite has the tendency to break up from the host grain in large size particles plugging the pore throats. Rahman et al. (1995) showed that the petrophysical properties of reservoirs containing illite depend significantly on the core preparation technique. Illite collapses upon air drying resulting in high porosity, high permeability and low capillary pressure. Illite rebounds, however, on contact with fresh water and projects across the pores, giving rise to low porosity, low permeability and high capillary pressure. Illite has also shown high susceptibility to migration into fresh water; it remains dispersed and is carried with the flowing fluid until the particles are trapped into pore constrictions. It is very difficult to distinguish clay migration damage from clay swelling damage. A steady, usually rapid decrease in permeability with decreasing salinity of the flowing liquids is generally a consequence of clay swelling; however, water sensitivity caused by particle migration will also appear in this case, but

[1] Return permeability or damage ratio (D.R.) was determined by comparison of initial (K_{is}) and final permeabilities (K_{fs}) in the stable state (D.R. = $100(K_{is} - K_{fs})/K_{is}$).

sometimes in a more irregular manner. Damage caused by particle plugging was detected by noting a temporary change (usually an increase) in permeability when fluid flow direction was reversed. In summary, porosity and permeability variations are function of several parameters, such as rock mineralogical composition, solids size, pressure, solution type and concentration.

Yan et al. (1996) reported that the optimal effect of bridging occurred when particle diameter is 1/2–2/3 of pore size. Regarding the reservoir heterogeneity with wide mineralogical composition, rock-fluid interaction affects permeability and porosity seriously. Amaefule et al. (1988) pointed out that five primary factors affect the mineralogical sensitivity of sedimentary formations: mineralogy and chemical composition, mineral abundance, mineral size, mineral morphology and mineral location. Mungan (1989) states that clay damage depends on the type and amount of exchangeable cations, such as K^+, Na^+, Ca^{++} and the layered structure existing in the clay minerals. The significance of damage during brine injection was observed to be a strong function of mineralogy and injection rate. The occurrence of a critical velocity, along with other observations, indicated that the primary damage mechanism was fines migration (Shenglai et al. 2008). A low pH may also contribute to less formation damage. At low pH, dissolution of silica and subsequently releasing of fines inside the formation is less.

6.2 Filtration

Filtration refers to the liquid phase of the drilling mud being forced into a permeable formation by differential pressure. During this process, the solid particles are filtered out, forming a filter cake. For filtration to occur, three conditions are required: a liquid or a liquid/solid slurry fluid and a permeable medium must be present and the fluid must undergo a higher pressure than the permeable medium. The knowledge of the filtration properties is very important in the design of drilling fluid formulation. Some works (Loeber, 1992; Li, 1996; Argillier et al., 1997; Benna et al., 1999 and 2001) have shown that the filtration across the cake depends on several parameters such as initial clay content, particle or aggregate association, water retention and permeability, as well as experimental conditions (pH, etc). Ferguson and Klotz (1954) showed that 70% to 90% of the total filtrate volume, flowing through permeable formations, occurred during mud circulation. During this dynamic filtration, the invasion radius reaches a value of 85%. A constant flow rate is reached when filtration forces, leading to the formation of a mud cake, are balanced by hydrodynamic forces, *i.e.* mud circulation that erodes the mud cake.

6.3 Damage mechanisms

The number of horizontally drilled wells has increased dramatically because they offer better contact with reservoir rocks, thus leading to higher production rates. Unfortunately, the larger drainage area contributes to longer exposure time for the drilling fluid. Consequently, fluid invasion cause severe damage, which would then have a considerable influence on productivity (Renard and Dupuy, 1991). Therefore a better understanding of damage mechanisms for various reservoir conditions can minimize the risks of horizontal well drilling and is still an important topic for research. Bishop (1997) summarized the seven mechanisms of formation damage previously reported by Bennion and Thomas (1994) and Civan (2001), as follows:

1. Fluid-fluid incompatibilities,

2. Rock-fluid incompatibilities,
3. Solids invasion,
4. Phase trapping/blocking,
5. Chemical adsorption/wettability alteration,
6. Fines migration,
7. Biological activity.

Some fields present several other formation damages due to salt, scales, or asphaltene deposits and/or clays and fines migration, arising from different sources, such as work-over and snubbing operations, perforations, cement filtrate invasion, reservoir pressure depletion and other pseudo-skin mechanisms, such as turbulence in production, partial penetration, completion problems...

Mineralogy and chemical composition, as well as the mineral by itself (size, morphology, abundance, and location), are considered as the primary factors affecting the mineralogical sensitivity of sedimentary formations (Amaefule et al., 1988). Moreover, Mungan (1989) stated that clay damage depends on the type and amount of their exchangeable cations.

Despite several studies and solutions proposed to reduce damage (Chang and Civan, 1997; Civan, 2000; Fisher et al., 2000; Parn-anurak and Engler, 2005), the problem is still not well understood. The main reason is the complex relationship existing between drilling fluid additives and rock reservoir composition. Both depend on porous medium structure and drilling conditions. Generally, sandstone pores are filled with a single phase or with multiphase fluids. It is widely recognized that experimental conditions i.e. overbalance pressure (Jilani et al. 2002 ; Al-Riyamy and Sharma, 2004), fluid type and composition (Longeron et al. 1995), filtration mode (Dalmazzone et al. 2006 ; van der Zwaag, 2006), solid concentration and pH (Baghdikian et al. 1989) can adversely affect permeability variations and thus cause different formation damage ratios. Some works were carried out to measure the drilling fluid invasion in Berea cores[2] at different overbalance pressures, keeping the other major influencing parameters constant, i.e. core permeability and nature of drilling fluid. Jilani et al. (2002) confirmed that return permeability increases when the overbalance pressure decreases but the invasion intensity starts to increase only after the overbalance pressure reaches a certain low 'critical' value. Overburden pressure can affect the results and parameters estimation seriously.

Due to the complexity of the foreign fluid/formation interaction and several other factors which affect the damage caused by fluid, the return permeability tests undertaken in the laboratory in a filtration cell (Figure 4) are generally the main tests used to explain damage mechanisms and differences noticed in fluid performances.

This test gives macroscopic information by measuring permeability variation, but cannot explain damage mechanisms. Other analyses are performed to study fluid rock interaction, such as additive retention and adsorption, wettability alteration, SEM (Scanning Electronic Microscopy) vizualization,…

---

[2] For the past 30 years, Berea Sandstone™ core samples have been widely recognized by the petroleum industry as the best stone rock for testing the efficiency of surfactants. It is a sedimentary rock whose grains are predominantly sand-sized and are composed of quartz sand held together by silica. The relatively high porosity and permeability of Berea Sandstone™ makes it a good reservoir rock.

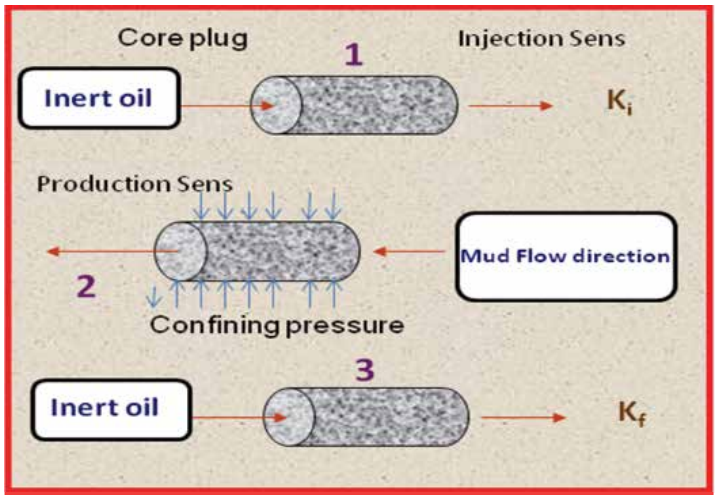


Fig. 4. Filtration cell for return permeability evaluation methodology

Many theoretical and experimental works have focused on the relationships between drilling fluid additive effect and return permeability. This last experimental method evaluates the relative permeability reduction and the effect of additive type and concentration with special scrutiny of solids and clay actions, leading to identify the most probable causes of damage during the selection process of candidate wells for matrix treatments. This also helps engineers to evaluate damage caused by clay swelling in reservoir conditions and to take the decision when to inject water in waterflooding operations.

6.4 Surfactants and wettability change

Drilling fluids contain various surfactants playing the roles of emulsifiers or wetting agents, and other specific additives, as mentioned above. Emulsion stability is often evaluated through rheological and filtration properties of the whole drilling fluid. Unfortunately, information on the nature and composition of emulsifiers is scarce. Acidic products from tall-oil, with an average chain length of 18 carbon atoms, other compounds deriving from tall-oil reactions with polyamines (Skalli et al., 2006), as well as alkanolamides, imidazolines and a variety of fatty acids and amines (Quintero, 2002) can act as emulsifiers. Similar surfactants are often used in traditional reverse emulsion systems (OBM) or in synthetic fluids (SOBM).

These surfactants cause mica and sandstone surfaces to become less water-wet (Skalli et al., 2006) and greatly reduce rock permeability (Yan and Sharma, 1989). Addition of surfactant to crude oil has less effect than sequential exposure of surfaces to crude oil and to surfactant solutions. Morever, the whole drilling muds have less effect on the wettability than their corresponding filtrates because of the presence of filter cakes which effectively prevent infiltration of large quantities of drilling fluid.

Among identified surfactants, which change the wettability of carbonate and sandstone rocks, let us mention i) fluorosilanes: only 1 wt.% concentration and 1 day aging period appear to be sufficient for altering wettability (Adibhatla et al., 2006) and ii) propyleneglycol (PG) (Audibert and Dalmazzone, 2006). The addition of PG to water-based drilling fluids can prevent the formation of in-situ water/oil emulsions and reduce the risk of water

blockage. Feng et al. (2009) showed that low-permeability and tight gas reservoirs still produce very strong water-blocking damage in the process of underbalanced drilling, and the lower the initial permeability, the bigger the water-blocking damage. However, the evaluation of water-blocking damage under underbalanced conditions is still in the exploratory stage.

7. Environmental considerations and waste management

Like polluted water and air, polluted soil can affect people health and environment through its action on surface waters (rain-out), underground waters and vegetation (phytotoxicity, bioaccumulation). The contamination may arise either through accidental discharge or uncontrolled industrial wastes. Undeniably, it constitutes one of the main environmental problems linked to the activities of oil and gas companies.

Since the early 90's, regulations do not authorize hydrocarbon losses and the closure of the site after drilling without treatment. Remediation technologies include dewatering, distillation, solvent extraction, cuttings reinjection, fixation, landfarming and (bio)remediation. All affect the economic and environmental acceptability of drilling operations.

To minimize the pollution due to OBM, numerous programs aim at reducing oil content according to regional and/or international standards. Recently, a new trend has gained increased support, namely the holistic approach to solve both drilling and waste problems (Getliff et al., 2000; Paulsen et al., 2001). Some concepts have been introduced to integrate economic and environmental considerations in drilling practices, such as Environmental Performance Indicators (EPI) (Jones et al., 1996) and Total Fluid Management (TFM) (Paulsen et al., 2002). Thus, much effort has been invested in exploring waste minimization opportunities (Greaves et al., 2001). In the 90's, drilling fluid companies devised new types of muds that used non-aqueous fluids (other than petroleum cuts). These fluids included linear paraffins, linear α-olefins, poly α-olefins, internal olefins and esters (Friedheim and Conn, 1996). Synthetic-based muds (SOBM), which have taken over an important niche in offshore drilling, share the desirable drilling properties of OBM but are free of polynuclear aromatic hydrocarbons and have lower toxicity, faster biodegradability and lower bioaccumulation potential. For these reasons, SOBM cuttings are less likely to cause adverse sea floor impacts than traditional oil-based cuttings (Drilling Waste Management Information System, 2004). The development of this new generation of synthetic fluids typically represents a compromise between environmental, economic, and performance considerations. This new approach, aimed at optimizing the design, delivery and management of wellsite fluids and wastes, exploits the natural grouping of all fluid-related products and services (Prutt and Hudson, 1998; Hudson and Nicholson, 1999; Hudson, 1999).

Currently, drilling fluid companies are developing fluid systems that are more amenable to biotreatment of the drilling wastes (Getliff et al., 2000; Growcock et al., 2002). It is likely that companies will continue to develop fluids with suitable drilling properties that contain fewer components or additives that would inhibit subsequent breakdown by earthworms or microbes. In some circumstances, mud components could serve as a soil supplement or horticultural aid.

Nevertheless, the loss of crude oil from producing wells, oil-based drilling fluids and refined petroleum products used in machinery operation and equipment remains the

primary source of contamination associated with drilling and production. The waste management assessment is geared towards solving waste problems through a logical developed process and using best practices and knowledge. However, one important question to answer is whether resource management adds any value to the exploration and production (E&P) business.

The philosophy behind the development of such fluids was not to design a system that merely posed a neutral or negligible impact on the environment, but rather one that would prove beneficial. Thus, the goal was to select the individual components of the fluid system, including the base fluid, emulsifiers, internal phase (salt and water), weight material and fluid-loss additives, to allow efficient drilling and generation of drill cuttings that can be used to actively enhance soil quality and subsequently support improved plant growth (Getliff et al., 2000). It is important to consider that the waste disposal method will function with the base fluid used in the continuous phase of the drilling fluid. For example, under the right environmental conditions, bacteria are very efficient at degrading many types of hydrocarbons. However those compounds that bacteria cannot readily degrade can delay the final remediation and close out of the site, thereby increasing the overall cost of the operation (Growcock et al., 2002). Alternatively, if the drilling fluid is optimized for its biodegradability by using a base fluid that does not contain any aromatic, cyclic or branched components, the treatment times can be significantly reduced, since there is no requirement to get rid of or reduce the humptane[3] fraction present in a diesel or mineral oil.

7.1 Waste treatment technologies: an overview

For oil companies, the great problem is to ensure an efficient environmental protection, avoiding over costs that might affect competitiveness. Therefore, the search for effective solutions at lower cost has a promising future. Currently, the treatment of waste from pits includes i) physical and chemical processes (removal of free phase, thermal desorption, excavation and disposal in landfill, deep injection in the wells, dehydration, incineration, neutralization, solidification and stabilization) and ii) biological processes (landfarming, biopiles, composting, phytoremediation and bioreactor).

Thermal techniques contribute significantly to the presence of heavy metals in aerosols. It is thus important to ascertain the quantities and chemical forms of the heavy metals that are emitted because their behavior strongly depends on the thermal and chemical environments.

Solidification/stabilization is currently applied on some mud pits and seems to be very effective. However, the use of solidification consists of a pollution transfer and/or containment without removing or even reducing the concentration of the initial soil pollution. Thus, solidification ensures the confinement of heavy metals and hydrocarbons and leaching, a simulation of rain wash, does not allow pollutant desorption. Losses can be due to a combination of factors including biodegradation, abiotic degradation, volatilization and migration (Khodja et al., 2005).

Biological treatments offer a suitable combination between economical issues and environmental protection. This should help operators to reduce drilling costs, while simultaneously increasing production and enhancing environment-oriented efforts. With a

[3]In Gas Chromatography, any unresolved components showing a broad "hump" (variously referred to as a "*humptane*" peak) above the baseline.

high potential for destroying environmental pollutants, bioremediation of crude oil-polluted soils (by degradation and detoxification) (Song et al., 1990) is becoming an increasingly important remedial option. The use of inexpensive equipment, the environmentally-friendly nature and simplicity of the process are some of its advantages over remedial alternatives such as physical and chemical treatments.

Numerous methods used for managing drilling wastes have been described in detail in the Drilling Waste Management Information System website (2004). They mainly include:

- **waste minimization**, which reduces volumes or impacts of wastes by minimizing the generation of drilling wastes thanks to special drilling techniques (*e.g.* directional drilling, smaller diameter holes, use of lower amount of fluid) or by using muds and additives with lower environmental impacts (e.g. SOBM or new drilling fluid systems or alternate weighting agents),
- **recycle/reuse**, *e.g.* mud recycle, roadspreading, reuse of cuttings for construction purposes, restoration of wetlands using cuttings or even use of oily cuttings as fuel,
- **other miscellaneous managing drilling waste methods** like disposal through onsite burial (pits, landfills), land application (landfarming, landspreading), bioremediation (composting, bioreactors, vermiculture), discharge to ocean, offsite disposal to commercial facilities, slurry injection, salt caverns or thermal treatments (incineration, thermal desorption). A combination of bioremediation with phytoremediation can afford better results for heavy metals. The selection of the treatment and the remediation technology of contaminated soils in the oil industry are then highly dependent on the drilling fluid composition but also on the environment regulations of the countries, geographic conditions, hydrogeology and climate of the drilled sites. Figure.5 shows drilling fluid waste sources and management methodology.

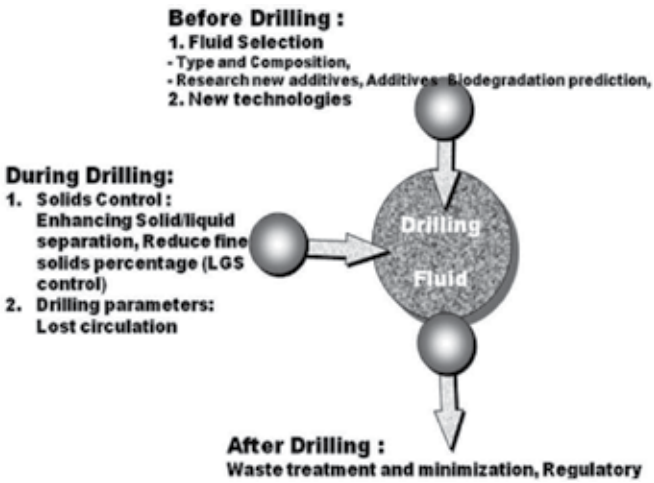


Fig. 5. Drilling fluid waste sources and management methodology

Life Cycle Assessment of drilling fluid

Using Life-Cycle Assessment (LCA) approach, recent work of Ghazi et al., (2008) evaluates the life cycle of all drilling muds, and compares four scenarios of treatment and disposal: thermal desorption, on-line and off-line stabilization/solidification and abandonment of reserve pit (without treatment). The preliminary results obtained show that LCA is a

relevant methodology to compare different scenarios of drilling mud treatments and to underline the step of the process which presents the major impact or damage factor.

7.2 Health effects

Some adverse health effects associated with drilling muds are i) irritation of the skin, eyes or alimentary mucosa caused by either low pH mud, surfactant or nuisance dust (de-aromatized hydrocarbons can enhance irritancy), ii) secondary irritation when prolonged and repeated contact (base oils/solvents) with skin will remove natural fats and oils to cause redness, drying and cracking, iii) respiratory irritation primarily from nuisance dusts, iv) inhalation effects such as acute central nervous system (CNS) depression when working with hydrocarbon solvents, especially at elevated temperatures. Solvents used in OBM are of low vapor pressure, and thus should not cause problems of CNS depression, although nausea and headache can occur (McDonald and Portier, 2003), v) possible sensitization to biocides and finally vi) possible carcinogenicity due to polycyclic aromatic hydrocarbons (PAH) and asbestos (Grieve, 1988).

8. Conclusions

Drilling process and drilling fluid formulation involve several parameters, which have to be taken into account, including:

- **Drilling cost in relation with the advanced technology**

The average cost for "conventional wells" (i.e., vertical wells drilled by using standard equipment) was about $75/ft (API, 1991) in the past two decades. This cost is related to the depth, type, and location of wells and also includes the costs of drilling-related services. A comparison with wells with similar depths and locations drilled a few years before indicates a decrease of this drilling cost, partially resulting from advances in technology.

- **Particle size as a source of benefit and damage**

In dispersion tests, our recommendation is to use, preferably, small size cuttings, which are in close contact with all additives used in drilling fluid systems. Moreover, when clays are fully exposed to the fluid, the aggregation effect is eliminated.

However, invasion of fine particles issued from drilling fluids (bentonite, barite, calcite,) also called 'mud invasion', or from geological sediments present in the formation, blocks the pores or at least narrows them down. This causes a decrease in the permeability of the formation thereby leading to a decrease in production rate. The economic consequences of this damage justify a thorough study of this problem in order to find ways to minimize its effects.

It has also been shown experimentally that WBM impair the formation permeability more significantly than OBM and polymer-based muds (Yan et al., 1996). The filtrate generated by WBM is more likely to cause physical interactions and chemical reactions with in situ reservoir fluid and rock, inducing severe damage.

Heavier OBM used to drill reservoir sections especially in undeveloped sectors where reservoir pressure is believed to be still high also leads to the main formation damage. This may be due to the particle invasion of the organophilic clays used as viscosifiers.

- **Waste management consideration**

Drilling fluid chemistry is quite complicated, and the effect of discharged mud into the environment is still not completely understood, despite a growing related research. For hydrocarbon decontamination, landfarming, a rapidly growing process, presents

satisfactory economic, scientific and environmental issues. In fact, biological techniques, cheaper than physical and chemical ones, prove to be very efficient in different soil types and ecosystems and present indeniable advantages such as no additional pollution, biodegradation either with autochthonous microorganisms or with added fertilizers or microorganism consortium. For heavy metals, a combination of bioremediation with phytoremediation can afford better results.

In conclusion, the authors assume that any technology vision and strategy should have strong anchor set with the business reality and should entirely adhere to industry core values. Worldwide exploration and production (E&P) operators are expecting integrated methods of preventing and minimizing existing waste management problems (i.e., waste generation, control and treatment). For maximum success, emphasis should be placed on assessment and planning rather than individual waste management products and technologies. The applied methodology should be focused on in-process responsive control and treatment methods rather than proposing an after-the-fact cure. The identified key of waste management needs falls into the general categories of engineering tools. Thus, the performance evaluation of all remediation techniques used so far can significantly help manager decision in choosing an appropriate waste treatment for each specific zone.

9. References

Aadnoy, B.S. (2003). Introduction to special issue on Borehole Stability, J. Petr. Sci. Eng., 38, 79-82

Abrams A. (1977). Mud design to minimize rock impairment due to particle invasion. *SPE 5713. JPT* May 586– 592.

Adibhatla, B.; Mohanty, K.K.; Berger, P. and Lee, C. (2006). Effect of surfactants on wettability of near-wellbore regions of gas reservoirs Journal of Petroleum Science and Engineering doi:10.1016/j.petrol.2006.03.026

Al-Riyamy, K. and Sharma, M.M. (2004). Filtration properties of oil-in- water emulsions containing solids. SPE paper 73769, pp. 164 -172

Amaefule, J.O.; Kersey, D.G.; Norman,D.L. and Shannon, P.M. (1988). Advances in Formation Damage Assessment and Control Strategies, CIM Paper No.88-39-65, Proceedings of the 39th Annual Technical Meeting of Petroleum Society of CIM and Canadian Gas Processors Association, Calgary, Alberta, June 12-16, 16 p

Anderson, E. E.; Cooper, G. A.; Maurer, W. C. and Westcott, P. A. (1991). An analysis of relative costs in drilling deep wells: SPE 22574, Proceedings of the Society of Petroleum Engineers 66th Annual Technical Conference and Exhibition, Dallas, Tex., Richardson, Tex., SPE, p.355-364.

American Petroleum Institute (1991). Joint Association Survey on 1990 Drilling Costs: Independent Petroleum Association of America, Mid-Continent Oil and Gas Association, 106 pp.

API 13A (1993). Specification for drilling fluid materials, American Petroleum Institute, Washington D.C., 15th ed., May

API (1997). Laboratory Testing of Drilling Fluids, Seventh Edition and ISO 10416: 2002 API 13I Supplement 2 - 01-June.

API (2003). Recommended Practice for Field Testing of water-based Drilling Fluids 13B-1, Third Edition, December, ANSI/API 13B-1/ISO 10414-1

Argillier, J.-F.; Audibert, A.; Janssen, M. and Demoulin, A. (1997). Performance of a new biodegradable ester-based lubricant for improving drilling operations with water based muds. Proceedings - International Symposium on Oilfield Chemistry, Houston (TX), Feb. 18-21, pp.539-549.

ASTM (2000). Standard test method for slake durability of shales and similar weak rocks. ASTM D4644-87, Annual Book of ASTM Standards, D18.

Audibert, A. and Dalmazzone, C. (2006). Surfactant system for water-based well fluids. Colloids and Surfaces A: Physico-chem. Eng. Aspects, 288, 113-120

Baghdikian, S.Y.; Sharma, M.M. and Handly, L.L. (1989). Flow of clay suspensions through porous media, SPE 16257 Reservoir Engineering, May 1989 pp.213-220

Bailey, L.; Keall, M.; Audibert, A. and Lecourtier, J. (1994). Effect of clay/ polymer interactions on shale stabilization during drilling. Langmuir 10, 1544–1549.

Beihoffer, T.W.; Dorrough, D.S. and Schmidt, D.D. (1988). The separation of electrolyte from rheological effects in studies of inhibition of shales with native moisture contents. SPE 18032. IADC/SPE Drilling Conference, Houston,(TX), 2-5 Oct..

Benna, M.; Khir-Ariguib, N.; Magnin, A. and Bergaya, F., (1999). Effect of pH on the rheological properties of purified sodium bentonite suspensions. J. Colloid Interface Sci. 218, 442–455

Benna, M.; Kbir-Ariguib, N.; Clinard, C. and Bergaya, F. (2001). Static filtration of purified sodium bentonite clay suspensions. Effect of clay content. Applied Clay Sci., 19, 103-120

Bennion, D. B. and Thomas, F.B. (1994). Underbalanced Drilling of Horizontal Wells: Does it really eliminate formation damage? *SPE 27352*, Formation Damage Control Symposium, February, Lafayette, Louisiana (1994)

Bergaya, F. ; Theng, B.K.G. and Lagaly, G. (Eds.) (2006). Handbook of Clay Science, vol.1 : Developments in Clay Science, 1246 pp., Elsevier Science, Amsterdam

Bishop, S. R. (1997). The Experimental Investigation of Formation Damage Due to the Induced Flocculation of Clays Within a Sandstone Pore Structure by a High Salinity Brine, SPE 38156 paper, presented at the 1997 SPE European Formation Damage Conference, The Hague, The Netherlands, June 2-3, pp. 123-143

Bland, R.G. (1991). Development of new water-based mud formulations. Chemicals in the Oil Industry: Developments and Applications. Spec. Publ.-R. Soc. Chem., Springer-Verlag, vol. 97, pp. 83– 98.

Bland, R.G. (1992). Water based glycol systems: acceptable substitute for oil-based muds. Oil Gas J., 54 (June).

Bland, R. G. (1994). Quality Criteria in Selecting Glycols as Alternatives to Oil - Based Drilling Fluid Systems, paper SPE 27141, HSE Conference, Jakarta, Indonesia

Bland, R.G.; Smith, G.L.; Eagark, P. and van Oort, E. (1995). Low salinity polyglycol water-based drilling fluids as alternatives to oil-based muds. Paper SPE/IADC 29378, IADC/SPE Drilling Conference, Amsterdam, Feb. 28 –March 2.

Bol, G.M.; Wong, S.W.; Davidson, C.J. and Woodland, D.C. (1992). Borehole stability in shales. Paper SPE 24975, SPE European Petroleum Conference, Cannes, Nov. 16–18.

Bradley, W.B. (1979). Failure of inclined boreholes. J. Energy Resour. Technol., Trans. AIME 102, 232–239.

Civan, F. (2000). Reservoir formation damage, Fundamentals, modelling, assessment and mitigation. Gulf Publishing company, Houston Texas

Civan F. (2001). Water sensitivity and swelling characteristics of petroleumbearing formations: Kinetics and correlation. SPE paper 67293

Chang, F. and Civan F. (1997). Practical model for chemically induced formation damage. Journal of Petroleum Science and Engineering. 1997. 17: 123-137

Chenevert, M.E. (1969). Shale Hydration Mechanics, SPE Paper 2401

Chenevert, M. E. (1970). Shale Alteration by Water Adsorption, Journal of Petroleum Technology, Sept., pp. 1141-1148.

Chenevert, M.E. and Pernot, V. (1998). Control of Shale Swelling Pressures Using Inhibitive Water-Bas Muds, SPE Paper 49263, 1998 SPE Annual Technical Conference and Exhibition in New Orleans, Louisiana, Sept. 27-30

Clark, R. K.; Scheuerman, R. F.; Rath, H. and Van Laar, H. G. (1976). Polyacrylamide-Potassium-ChlorideMud for Drilling Water-Sensitive Shales, paper SPE 5514, Journal of Petroleum Technology, June, pp. 719–727.

Chilingarian, G. V. and Vorabutr, P. (1983). Drilling and Drilling Fluids, Elsevier Scientific, Amsterdam

Clark, D.E. and Saddok, B. (1993). Aluminium Chemistry Provides increased shale stability with Environmental acceptability. SPE 25321 paper presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition, Singapore, Feb. 8-10

Cliffe, S.; Dolan, B. and Reid, P.I. (1995). Mechanism of shale inhibition by polyols in water-based drilling fluids. Paper SPE 28960 presented at the SPE International Symposium on Oilfield Chemistry, San Antonio, Feb.

Cole, M.G. (1994). Assessment and Remediation of Petroleum Contaminated Sites, CRC Press, Boca Raton, FL, 63

Cook, J.M.; Goldsmith, G.; Geehan, T.M.; Audibert, A.M.; Bieber, M.T. and Lecourtier, J. (1993). Mud/Shale interaction: model wellbore studies using X-ray tomography. Drilling Conference, paper SPE/IADC 25729, Amsterdam, Feb. 23-25,

Coussot, P.; Bertrand, F. and Herzhaft, B. (2004). Rheological Behavior of Drilling Muds, Characterization Using MRI Visualization. Oil & Gas Sci. Technol. – Rev. IFP, 59(1,) 23-29

Dalmazzone, C.; Audibert-Hayett, A.; Quintero, L.; Jones, T.; Dewatinnes, C. and Jansen, M. (2006). Optimizing filtrate design to minimize in-situ and wellbore damage to water-wet reservoirs during drill-in. SPE Prod. 66-73, Feb.

Darley, H.C.H. (1969). A laboratory investigation of borehole stability. J. Petrol. Technol. July, 883-892.

Darley, H.C.H. and Gray, G.R. (1988). Composition and properties of drilling and completion fluids. 5th Ed., Gulf Professional Publishing, Houston, TX, 630p

Ding, R.; Qiu, Z. and Li, J. (1996). Soluble-silicate mud additives inhibit unstable clays. Oil and Gas J., 66– 68, Avil

Downs, J.D.; van Oort, E.; Redman, D.; Ripley, D. and Rothmann, B. (1993). TAME—a new concept in water-based drilling fluids. Paper SPE 26699 presented at the Offshore Europe Conference, Aberdeen, September 7 –10

Drilling and Excavation Technologies for the Future Committee on Advanced Drilling Technologies (1994). National Research Council, ISBN: 0-309-57320-3, 176 pages, 6 x 9, National Academy of Sciences

Drilling Waste Management Information System website (2004). launched in 2004. http://web.ead.anl.gov/dwm.

Durant, C. ; Forsans, T. ; Ruffet, C. ; Onaisi, A. and Audibert, A. (1995). Influence of clays on borehole stability: Part one (Occurrence of drilling problems physico-chemical description of clays and of their interaction with fluids). Revue de l'IFP, Vol. 50, N°2, Mars-Avril

Feng, Z.; Hongming, T.; Yingfeng, M.; Gao, L. and Xijin, X. (2009). Damage evaluation for water-based underbalanced drilling in low-permeability and tight sandstone gas reservoirs Petrol. Explor. Develop., 36(1): 113-119.

Ferguson, C.K. and Klotz, J.A. (1954). Filtration from mud during drilling. Trans. AIME, 201: 29-42

Fjaer, E. (1999). Static and dynamic moduli of weak sandstones. Proceedings of the 17th US Rock Mechanics Symposium, Vail, Colorado, 6-9 June.

Friedheim, J.E. and Conn, H.L. (1996). Second generation synthetic fluids in the North Sea: are they better? Paper presented at the IADC/SPE Drilling Conference, New Orleans, SPE 35061

Friedheim, J.; Toups, B. and van Oort, E. (1999). Drilling Faster with Water-Based Muds. AADE Houston – Annual Forum – Improvements in Drilling Fluid Technology, Houston (TX), March 30-31.

Fritz, S.J. and Marine, I.W. (1983). Experimental Support for a Predictive Osmotic Model of Clay Membranes, Geochim. Cosmochim., Acta 47, 1515-1522

Ghazi, M.; Quaranta, G.; Duplay, J. and Khodja, M. (2008). Life-Cycle Assessment (LCA) of drilling mud in arid area. Evaluation of specific fate factors of toxic emissions to groundwater. First results''. SPE-111646. This paper was prepared for presentation at the 2008 SPE International Conference on Health, Safety, and Environment in Oil and Gas Exploration and Production held in Nice, France, 15–17 April

Getliff, J.M.; Bradbury, A.J.; Sawdon, C.A.; Candler, J.E. and Loklingholm, G. (2000). Can advances in drilling fluid design further reduce the environmental effects of water and organic-phase drilling fluids?, paper SPE 61040 presented at the Fifth SPE International Conference on Health, Safety and Environment, Stavanger, Norway, 26-28 Juin

Greaves, C.; Rojas, J.C. and Chambers, B. (2001). Field Application of "Total Fluid Management" of Drilling Fluids and Associated Wastes, paper SPE 66552 presented at the SPE/EPA/DOE Exploration and Production Environmental Conference, San Antonio, Texas, 26-28 February

Grieve, A. (1988). Toxicity of drilling mud. Journal of Occupational Health 40(12):736–739

Grolimund, D.; Barmettler, K. and Borkovec, M. (2001). Water Resource Res. 37:571

Growcock, F.B.; Curtis, G.W.; Hoxha, B.; Brooks, W.S. and Candler, J.E. (2002). Designing Invert Drilling Fluids to Yield Environmentally Friendly Drilled Cuttings, IADC/SPE 74474, IADC/SPE Drilling Conference, Dallas, TX, February 26-28.

Hale, A.H. Blytas, G.C. and Dewan, A.K.R. (1989). U.K. Patent application N°. 2, 216, 574

Hale, A.H. and Mody, F.K. (1992). Experimental investigation of the influence of chemical potential on wellbore stability. IADC/SPE 23885.

Hale, A. H. and Mody, F. K. (1996). Experimental Investigation of the Influence of Chemical Potential on Wellbore Stability, IADC/SPE Paper 23885, Presented at the SPE/IADC Drilling Conference in New Orleans, Louisiana, February 18-21

Haus, F.; Boissel, O. and Junter, G-A. (2003). Multiple regression modelling of mineral base oil biodegradability based on their physical properties and overall chemical composition. Chemosphere 50 939–948

Himel, C.M. and Lee, E.G. (1951). Drilling fluids and methods of using same. U.S. Pattent N°2,570,947

Hongwei, Y.; Zhanpeng, J. and Shaoqi, S. (2004). Science of the total Environment 322, 209-219

Horsud, P.; Bostrom, B.; Sonstebo, E. F. and Holt, R. M. (1998). Interaction Between Shale and Water-Based Drilling Fluids: Laboratory Exposure Tests Give New Insight Into Mechanisms and Field Consequences of KCl Contents, Presented at the SPE Annual Technical Conference and Exhibition in New Orleans, LA, Sept.

Hudson, C. (1999). Evaluation of drilling rig fluids handling systems: an integrated fluids management approach, Offshore Magazine, September

Hudson, C. and Nicholson, S. (1999). Integrated fluids approach cuts waste, costs in Texas Wildlife refuge, Petroleum Eng. International, pp 37-41, March

Jilani. S.Z,; Menouar, H.; Al-Majed, A.A. and Khan, M.A. (2002). Effect of overbalance pressure on formation damage, J. Petr. Sci. Eng. 36, 97– 109

Jones, M. G.; Hartog, J. J. and Sykes, R.M. (1996). Environmental Performance Indicators - the Line and Management Tool. SPE 35953. lnternational Conference on Health, Safety and Environment, New Orleans, LA, 9.12 June

Kelly, J. (1968). Drilling problem shales 1 : Classification simplifies mud selection. Oil and gas J. (3 June), 66(23),67-70

Khodja, M.; Hafid S. and Canselier, J. P. (2005). A Diagnostic of the Treatment and Disposal of Oil Well Drilling Waste, WasteEng 05, Albi, France, Proceedings on CD-Rom, May

Khodja, M (2008). Drilling Fluid : Performance Study and Environmental Consideration, Thesis National Polytechnic Institute, Toulouse, France

Khodja, M; Canselier, J.P.; Bergaya, F.; Fourar, K.; Khodja-Saber, M.; Cohaut, N. and Benmounah, A. (2010). Shale problems and water-based drilling fluid optimization in Hassi Messaoud Algerian oil field. Appl. Clay Sci., 49, 383–393

Lamberti CMC (1999). Polymers and performance chemicals, Laboratory Test - PAC used in drilling fluid formulation, Albizzate, Italy.

Leung, P.K. and Steig, R.P. (1992). Dielectric constant measurement: A new, rapid method to characterize shale at the wellsite. IADC/SPE Drilling conference, 18-21 Feb., New Orleans, LA, USA

Li, Y. (1996). Filtration statique et dynamique de différents systèmes argile, électrolytes, polymères. Thesis University of Orléans, France

Likos, W.J.; Loehr, J. E. and Akunuri, K. (2004). Engineering Evaluation of Polymer-Based Drilling Fluids for Applications in Missouri Shale" University of Missouri-Columbia. July

Loeber, L. (1992). Study of clay mud cake structure formed during drilling. Thesis, University of Orléans, France.

Lomba, R. F., Chenevert, M. E. and Sharma, M. M.(2000). The ion-selective membrane behavior of native shales", J. Petr. Sci. Eng., pp. 9-23

Longeron, D.; Argillier, J-F. and Audibert, A. (1995). An integrated experimental approach for evaluating formation damage due to drilling and completion fluids. SPE 30089, European formation damage conference, The hague, the Netherlands, 15-16 May

Low, F.P. and Anderson, D.M. (1958). Osmotic pressure equation for determining thermodynamic properties of soil water. Soil sci. 86, 251-258

Lummus, J.L. and Azar, J.J.(1986). Drilling Fluids Optimization : A Practical Field Approach, PennWell Books, Tulsa (OK), pp. 3-5

Lutton, R. J. (1977). Slaking Indexes for Design (Report FHWA-RD-77-1), Design and Construction of Compacted Shale Embankments, US Department of Transportation, Vol.3. 88 pp, Federal Highway Administration, Washington DC.

McDonald, Jason A. and Portier, Ralph J. Feasibility studies on *in-situ* biological treatment of drilling muds at an abandoned site in Sicily, J Chem Technol Biotechnol. 78:709–716 (online: 2003) DOI: 10.1002/jctb.847

Mody, F. K. and Hale, A. H. (1993). Borehole-Stability Model To Couple the Mechanics and Chemistry of Drilling Fluid/Shale Interactions, J. Petr. Technol., Nov., pp. 1093-1101.

Mungan, N. (1989). Discussion of An Overview of Formation Damage, J. Petr. Technol., 41(11) 1224

Muniz, E.S.; Fantoura, S.A.B. and Lomba, R.F.T. (2004). Development of equipment and testing methodology to evaluate rock-drilling fluid interaction. GulfRocks04, The 6th North America rock mechanics symposium (NARMS), Houston, Texas, paper 599, 8p.

Nair, N. G. (2004). Asphaltic shale coating agent, Master of Science in Engineering presented to the Faculty of the Graduate School of The University of Texas at Austin, Houston, p.53

Patel, A.; Stamatakis, E.; Friedheim, J.E. and Davis, E. (2001). Highly inhibitive water-based fluid system provides superior chemical stabilization of reactive shale formations. Drilling Technology, American Association of Drilling Engineers, AADE National Drilling Technical Conference, AADE 01-NC-HO-55, Houston,(TX), March 27-29.

Patel, A., Stamatakis, E., Young, S. and Stiff, C. (2002). Designing for the future - A review of the design, development and inhibitive water-based drilling fluid. Drilling and Completion Fluids and Waste Management, Houston (TX), April 2-3.

Paulsen, J.E.; Saasen, A.; Jensen, B. and Grinrød, M. (2001). Key Environmental Indicators in Drilling Operations, paper SPE 71839 presented at the Offshore Europe Conference held in Aberdeen, Sept. 4-7

Paulsen, J. E.; Saasen, A.; Jensen, B.; Thore Eia, J. and Helmichsen, P. (2002). Environmental Advances in Drilling Fluid Operations Applying a Total Fluid Management Concept. Paper presented at the AADE 2002 Technology Conference. Drilling and Completion Fluids and Waste Management, held at the Radisson Astrodome. Houston, Texas, April 2 - 3,

Pernot, V. F. (1999). Troublesome Shale Control Using Inhibitive Water-Base Muds, Thesis Presented to the Faculty of the Graduate School of the University of Texas at Austin, May

Perricone, A.C.; Clapper, D.K. and Enright, D.P. (1998). U.S.Patent N°.4, 830, 765

Prutt, J. and Hudson, C. (1998). Integrated approach optimizes results, American Oil and Gas Reporter, August, pp. 86-91

Quintero, L. (2002). An Overview of Surfactants Applications in Drilling Fluids for the Petroleum Industry, J. Disp. Sci. Technol., 23 (1-3), 393-404

Rahman, S.S.; Rahman, M.M. and Khan. F.A. (1995). Response of low-permeability, illitic sandstone to drilling and completion fluids, J. Petr. Sci. Eng., 12, 309-322

Reid, P.I.; Elliott, G.P.; Minton, R.C.; Chambers, B.D. and Burt, D.A. (1993). Reduced environmental impact and improved drilling performance with water-based muds containing glycols. Paper SPE 25989 presented at the SPE/EPA Exploration and Production Environmental Conference, San Antonio, March 7 – 10.

Renard, G. and Dupuy, J.G. (1991). Formation damage effects on horizontal-well flow efficiency. J. Pet. Technol., 43 7 786 – 869, July

Retz, R.H.; Friedheim, J.E.; Lee, L.J. and Welch, O.O. (1991). An environmentally acceptable and field-practical, cationic polymer mud system. Paper SPE 23064 presented at the Offshore Europe Conference 3– 6 Sept..

Roehl, E.A. and Hackett, J.L. (1982). A laboratory technique for screening shale swelling inhibitors. SPE paper 11117, 57th Annual Fall Meeting, Society of Petroleum Engineers, New Orleans (LA), 26 Sept..

Rosana, F.T.; Lomba, M.E.; Chenevert, M. and Sharma, M. (2000). The role of osmotic effects in fluid flow through shales. J.Petr. Sci. Eng., 25 ,25–35

Roy, SB. and Dzombak, DA. (1996). Colloids Surf A Physicochem Eng Asp 107:245

Saddok, B.; Clapper, Dennis K.; Parigot, P. and Degouy, D. (1997). Oil Field Applications of Aluminium Chemistry and Experience With Aluminium-Based Drilling Fluid Additive. SPE 37268 paper presented at the International Symposium on Oilfield Chemistry, Houston, TX 18-21 Feb.

Saiers, JE. and Hornberger, GM. (1999). Water Resour. 35:1713

Salisbury, D.P. and Deem, C.K. (1990). Tests show how oil muds increase shale stability. World oil, Oct.p.57-65

Schlemmer, R.; Friedheim, J. E.; Growcock, F. B.; Headley, J. A.; Polnaszek, S. C. and Bloys, J. B. (2002). Membrane Efficiency in Shale – An Empirical Evaluation of Drilling Fluid Chemistries and Implications of Fluid Design, IADC/SPE Paper 74557, Presented at IADC/SPE Drilling Conference in Dallas, Texas, Feb. 26-28

Shenglai, Y.; Zhichao, S.; Wenhui, L.; Zhixue, S.; Ming W. and Jianwei Z. (2008). Evaluation and prevention of formation damage in offshore sandstone reservoirs in China Pet.Sci.5:340-347

Simpson, J.P.; Walker, T.O. and Jiang, G.Z. (1994). Environmentally acceptable water-base mud can prevent shale hydration and maintain borehole stability. Paper IADC/SPE 27496 presented at the IADC/SPE Drilling Conference, Dallas, TX, 15–18 Feb.

Skalli, L.; Buckley, J.S.; Zhang, Y. and Morrow, N.R. (2006). Surface and core wetting effects of surfactants in oil-based drilling fluids doi:10.1016/j.petrol.2006.03.012 J. Petr. Sci. Eng., 52, 253-260

Song, H.G.; Wang, G.X. and Bartha, R. (1990). Bioremediation potential of terrestrial fuel spills, Appl. Environ. Microbiol 56 : 641-651

Sorheim, R.; Amundsen, C.E.; Kristiansen, R. and Paulsen, J.E. (2000). Oily Drill Cuttings – From Waste to Resource, SPE paper no. 61372, presented at the Fifth SPE International Conference on Health, Safety and Environment, Stavanger, Norway, 26-28 June

Stamatakis, E.; Thaemlitz, C.J.; Coffin, G. and Reid, W. (1995). A new generation of shale inhibitors for water-based muds, SPE/IADC 29406 Drilling Conference

Steiger, R. P. (1993). Advanced Triaxial Swelling Tests on Preserved Shale Cores", Presented at the 54th U.S Symposium on Rock Mechanics, June 27-30

Swartz, CH. and Gschwend, PM. (1998). Environ Sci Technol. 32:1779

Swartz, CH. and Gschwend, PM. (1999). Water Resour Res. 35:2213

Tare, U.A. and Mody, F.K. (2000). Stabilizing boreholes while drilling reactive shales with silicate-based drilling fluids. Drilling Contractor, May/June, 42-44.

Twynam, A.J.; Caldwell, P.A. and Meads, K. (1994). Glycol-enhanced water-based muds: case history to demonstrate improved drilling efficiency in tectonically stresses shales. Paper IADC/SPE 27451 presented at the IADC/SPE Drilling Conference, Dallas, TX, Feb. 15–18

van der Zwaag, C.H. (2006). Benchmarking the formation damage of drilling fluids. SPE Prod. Oper 40-50,

van Oort, E. (1994). A novel technique for the investigation of drilling fluid induced borehole instability in shales. Paper SPE/ ISRM 28064 presented at the SPE/ISRM Conference on Rock Mechanics in Petroleum Engineering, Delft, Aug. 29–31.

van Oort, E.; Ripley, D.; Ward, I.; Chapman, J.W.; Williamson, R. and Aston, M. (1999). Silicate-based drilling fluids: competent, cost-effective and benign solutions to wellbore stability problems. SPE paper 35059, SPE/IADC Drilling Conference, New Orleans (LA), March 12-15.

van Oort, E. (2003). On the physical and chemical stability of shales. J. Petr. Sci. Eng., 38, 213–235

van Oort, E.; Hale, A. H. and Mody, F. K. (1996a). Transport in Shales and the Design of Improved Water-Based Shale Drilling Fluids, SPE Drilling and Completion, Sept., pp. 137-146.

van Oort, E.; Hale, A.H., van Oort, E.; Ripley, D.; Ward, I.; Chapman, J.W.; Wiliamson, R. and Aston, M. (1996b). Silicate-based drilling fluids: competent, cost-effective and benign solutions to wellbore stability problems. Paper SPE 35059 presented at the IADC/SPE Drilling Conference, New Orleans, LA, March 12– 15.

Ward, I. and Williamson, R. (1996). Silicate water based muds—a significant advance in water based drilling fluid technology. Paper presented at the IBC Conference on the prevention of oil discharge from drilling operations, Aberdeen, June 18- 19.

Ward I.; Chapman, J.W. and Williamson, R. (1997). Silicate-based muds: chemical optimization based on field experience. SPE paper 55054, SPE International Symposium on Oilfield Chemistry, Houston (TX), Feb. 18-21, pp. 551-560.

Wilcox, R.D.; Fisk, J.V. Jr. and Corbett, G.E., (1987). Filtration method characterizes dispersive properties of shales. SPE Paper 13162, 59th Annual Technical Conference and Exhibition, Houston (TX), SPE Drilling Engineering Journal, June, 2(2), 149-158.

Wood, L.E. and Deo, P. (1975). A suggested system for classifying shale materials for embankments. Bulletin Assoc. Eng. Geologists 12 (1), 39-55.

Yan, J. and Sharma, M.M. (1989). Wettability alteration and restoration for cores contaminated with oil-based muds, J. Petr. Sci. Eng., 2, 63-76,

Yan, J.; Jiang, G. and Wang, F., (1996). SPE-37123. 2nd International Conference on Horizontal Well Technology, Calgary, Alberta Canada, 18– 20 Nov. 739–747

Young, S.Y. and Maas T., (2001). Novel polymer chemistry increases shale stability. Drilling Technology, American Association of Drilling Engineers, AADE National Drilling Technical Conference, AADE 01-NC-HO-41, Houston (TX), March 27-29.

Zamora, M. ; Lai, D.T. and Dzialowski, A.K. (1990). Innovative Devices for testing Drilling Muds, SPE Drilling Eng. J., 5(1), March, 11-18.

Zhang, J.; Chenevert, M. M.; Talal, A. and Sharma, M. M., (2004). A new gravimetric-swelling test for evaluating water and ion uptake of shales. SPE 89831, SPE Annual Technical Conference and Exhibition, Houston (TX), Sept. 26-29.

Zhang, J.; Clark, D.E.; Al Bazali, T.M.; Chenevert, T.; Sharma, M.M.; Rojas, J.C. and Seehong, O. (2006). Ion movement and laboratory technique to control wellbore stability. AADE-06-DF-HO-37. Fluid conference, Houston TX, April 11-12

Zhanpeng, J.; Hongwei, Y.; Lixin, S. and Shaoqi, S. (2002). Integrated assessment for aerobic biodegradability of organic substances. Chemosphere 48 ,133-138

14

The Advanced Technologies Development Trends for the Raw Material Extraction and Treatment Area

Ján Spišak, PhD. and Miroslav Zelko, PhD.
Development and realisation workplace of raw materials gaining and treatment, Faculty of Mining, Ecology, The Technical University of Košice Slovakia

1. Introduction

Mining organizations operating in today's market face many complex challenges. The continuing globalization of the mining industry means both increased demand and increased competition. This increased demand and strong commodity prices mean that improving productivity of processes is essential, while at the same time needing to control costs and maintain an effective cost profile for the future. To be effective and to keep productivity at optimum levels, the organization's logistics function is crucial. Ensuring continuity of the in-bound supply chain, maintaining service levels, maximize efficiency of outbound logistics, reducing inventory and increasing inventory turns are all essential to maintaining productivity and minimizing costs. In addition to the productivity and logistical challenges, the mining industry is increasingly the focus of public attention with regard to its health and safety and environmental performance. In order to survive and thrive in this dynamic environment, mining companies need powerful and flexible strategies and tools to enhance the way they operate.

The **EU minerals industry** provides vital inputs to Europe's economy and social well-being – it is not only an important supplier to the EU economy, it is also a world leader supplier of services, technology, engineering, consultancy, finance and equipment. Because of the high environmental standards and the often challenging geological conditions prevailing in Europe, European extractive technology has a leading position and holds about 50% of the relevant world market. To sustain this position and to ensure next development sustainability an intensive research and development on advanced technologies and high-tech products based on raw materials is needed. In 2008 the European Commission published its new strategic initiative "The raw materials initiative — meeting our critical needs for growth and jobs in Europe". Raw materials are essential for the sustainable functioning of all societies, so this platform could be a decisive priority for next EU development.

The requirement to realize a society development in intension of permanent technological sustainability resulted in many concepts and strategic documents or platforms, among others to **The European Technology Platform on Sustainable Mineral Resources** (ETP SMR). ETP-SMR unites hundreds of stakeholders from industry, the research community, public authorities, the financial community, regulators, consumers and civil society around

the major technological challenges to the sector, in order to jointly develop a common vision. The platform will contribute to strengthening one of the fundamental pillars of the European economy and society: the European minerals industries. These include oil, gas, coal, metal ores, industrial minerals, ornamental stones, aggregates, smelters as well as technology suppliers and engineering companies. The ETP SMR has the following objectives: securing the future supply of/access to European raw materials; supporting the revival of exploration of Europe's mineral potential; developing innovative and sustainable production technologies; implementing best practices; reuse, recovery and recycling as well as new product applications; creating European added value through RTD-based technology leadership, education and training.

2. The importance and overview of advanced technologies

Currently there has been no shared understanding within the EU on exactly what should be considered as key advanced technologies. There is no coherent strategy on European level on how these technologies can be better brought to industrial deployment at a European level. Generally, the advanced technologies are knowledge intensive and associated with high R&D intensity, rapid innovation cycles, high capital expenditure and highly-skilled employment. They enable process, goods and service innovation throughout the economy and are of systemic relevance. They are multidisciplinary, cutting across many technology areas with a trend towards convergence and integration.
The advanced technologies and materials are the basis of the future priority to improve European industrial competitiveness. The top-ranking technologies are the part of ***advanced manufacturing*** *systems* leading to improvements in terms of new product properties, production speed, cost, energy and materials consumption, operating precision, waste and pollution management. The advanced technologies in mining industry will be based on marketable knowledge-based systems and the related services (e.g. simulation of automated robotics, extraction and finishing lines). Advanced technologies can be applied in all manufacturing industries and form an important element in the supply chain of many high value manufacturing businesses. They make up some 10.5% of EU industrial productions and provide some 2.2 million jobs and account for 19% of EU exports and over 40% of EU private sector R&D expenditure (ETP-SMR, 2009).

2.1 World-wide key concepts for the raw material extraction and treatment area

a. CSIR (RSA) „FutureMine" project - continuation of „DEEPMINE" in the area of occupational health and ergonomics, issues for deep mining mechanisation, automation, communication and sensors, http://www.csir.co.za/index.html. The Council for Scientific and Industrial Research (CSIR) in South Africa is one of the leading scientific and technology research, development and implementation organisations in Africa. It undertakes directed research and development for socio-economic growth.
b. CSIRO Exploration & Mining (Australia), main issues: Sustainable Mining Systems, Mining Automation, Mining Geoscience, Next Generation Mineral Mapping, http://www.em.csiro.au/about/about.htm. CSIRO, the Commonwealth Scientific and Industrial Research Organisation, is Australia's national science agency and one of the largest and most diverse research agencies in the world.
c. DMRC (Canada) Deep Mining Research Consortium – current DMRC projects are defined by the current challenges of deep mining in Canada,

http://www.deepminingresearch.org/Projects.htm. The DMRC provides a forum for members to fund research to improve or develop new technologies for mining at depth. The DMRC membership includes seven mining companies, the City of Greater Sudbury and CANMET-MMSL.

2.2 European concepts for the raw material extraction and treatment area

1. **Intelligent Mine** - the concept of Helsinki University of Technology.

According to this concept, it is a mine that monitors its entire operation in real time, with each process feeding relevant data to the successor process for action, as well as to the predecessor process for feedback. **(**Mining Congress, Katowice, Sept. 2008). The concept is based on the Production Management System and Data visualization and real-time production control system (Särkkä, 2008).

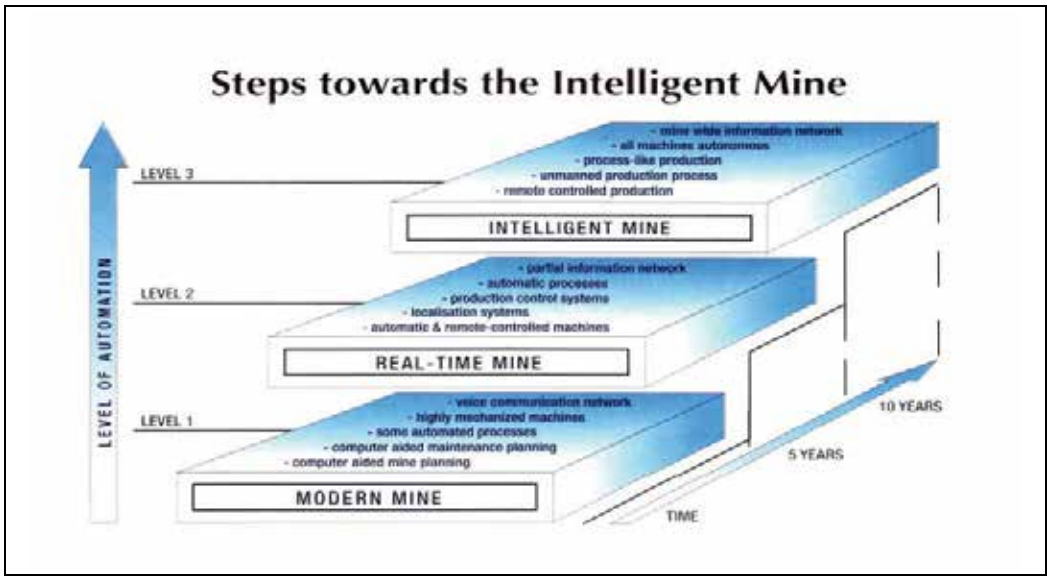


Fig. 1. Intelligent Mine concept

Production Management System (PMS) - is a real time control and monitoring system concerning the whole production chain of a mine starting from geological data ending to delivered concentrates. In the system all the needed applications should be integrated to use the one and the same database. The system should allow manage all the information needed in daily operation. Usually such a system consists of:

- Mine planning
- Production monitoring
- Production management
- Maintenance management
- Real-time condition monitoring and fault diagnostics
- Process automation
- Remote control tele operation

The production control part includes both the work planning and the real-time data collection from the production chain as the feedback to the scheduling. In addition to

conventional daily or monthly production reports the key performance indicators can be followed in any timescale desired. The condition control part supports the maintenance work planning updating the information associated with the production equipment and adapting services into the production plan. Furthermore such a system gives the tools for the general management of maintenance.

Data visualization and real-time production control

The structure of the PMS is usually open. The data management system manages the visualizations, simulations, optimizations and production processes, resources location and automated equipments on the other side. The browser-type user interface is generally tailored according to expectations of the mine personnel. The user interface contains very easy-to-use search, assorting and linking features. All the information saved in PMS, dealing with geology, mining, milling, maintenance, cost control all the way to the system administration, should be managed with the same user interface.

2. **Mine of the Future (MIFU)**

The Swedish Initiative – Mine of the Future for years of 2009 – 2010, leader: Nordic Rock Tech Centre AB (RTC) established a consortium for the conceptual study "Mine of the Future" (MIFU) to develop a common vision for future deep mining (depth 1,500-2,000 m). The MIFU main tasks within a Strategic Research, Development and Innovation Agenda 2011- 2020 is research and development of new production processes and technologies and to conceptualise new objectives: the Attractive Workplace, lean and Green Mining and Mineral Processing and new Production Processes and Technologies.

MIFU is the result of growing social order, which emerged from the following facts:

- Extraction is realized from deeper mines and with lower ore grade;
- Stiffer environmental regulations (less energy, less CO_2, less water – necessary);
- Difficult to attract young people for a life in remote locations;
- Challenging safety standards in deep mine conditions;
- Market need for metals will increase;
- Big mines get bigger, smaller mines get smaller and more selective;
- Only the big mining companies have the resources and capabilities to develop and operate on global scale;
- Challenges and the changes are so large and numerous that a comprehensive international cooperation is needed both within and outside the industry in order to succeed.

3. **Intelligent Production Systems for a Sustainable Supply and Use of Mineral Resources (EU-IPSUM)**

Leader is prof. Nicolai Martens, Ludger Rattmann, Institute for Mining Engineering RWTH Aachen University. Intelligent Production Systems (EU-IPSUM) idea consists of:

- **Intelligent mass movement** (transport systems, logistics, process control and automation, ...)
- **Innovative rock breaking technologies** (cutting technologies, SMART IT-controlled blasting systems)
- **Intelligent mineral processing** (near to face processing, multi mineral processing, bio-processes)
- **Intelligent mobile production systems** (keyhole mining, sustainable industrial small scale mining)
- **New geo-resources** (geothermal energy, mining oilfields, water, ...)

- **Information Technologies** (sensors, RFIDs, process control, ...)

Common features of next advanced technology built concept are:

- **Production Management System –** mine planning, production monitoring, maintenance monitoring
- **Real-time condition monitoring and fault diagnostics** – localization systems,
- **Informatization and digitalization** - data visualization, simulation and optimization, virtual reality principles
- **Real-time production control –** technological logistics, RFID systems, sensors
- **Process automation –** SMART factory principles, automatic and remote-controlled machines
- **Intelligent mine** – unmanned production processes, autonomous machines, mine-wide information network

3. Innovative megatrends – basis for the design of advanced technologies

Each period is characterized by basic directions of development, which are determined by existing terms and conditions. Its early tracing allows to obtain a better position of the existing industry resp. companies in this sector through the increased productivity, quality production and cost reduction. In these directions should be developed main innovative activities in the future. In the present, the main lines of development are in the field of technology, logistics and information. In terms of exploitation and processing of raw materials we consider these trends for the most important:

- informatization and digitalization,
- using virtual reality,
- technological logistics,
- advanced control of processes,
- process approach, modeling and simulation,
- results from the research and development of technologies for the raw material extraction and treatment,
- customer oriented production – pull system of material flows,
- ecology and safety.

3.1 The informatization and digitalization

Basis for the **informatization** is process approach and its main aim is to obtain adequate information about realized resp. designed processes. The main source of information about processes are operational measurments, physical and mathematical models. Operational measurments provide empirical information. For obtaining a categorial information we need model experiments. Mathematical models have unique advantage. Their main attibutes must be adequacy and speed of medelling processes. One way how to achieve these requirements is the creation of models based on physical principles and creation of the simplified – replacing models. Simplified patterns must be created for each specific problem. To make creation of models more effective it is useful to apply an appropriate support system. Creation of such systems is very actual. We can consider a development of process models as a dominant factor of their computerization. A second key factor is process analysis, which task is to obtain relevant information needed for the further knowledge.

Closely related with the informatization is also internet and using of advanced communication and identification systems particulary in the logistic processes. One example of advanced technologies for the identification is system RFID. Its use in the extraction and processing of raw materials allows a precise identification of individual doses and following automated process control. For example For example. after arriving into the supply bin, chip RFID gives an impulse, by which provides us with an information about the quality, size of the dose, planned mode of processing etc.

Digitalization enables to increase the quality and at the same time expedite all works related to production preparation, primary production and following services in the total product life cycle. The concept of digitalization represents through its ICT integrated environment in which the reality is substitute by virtual computer models. These virtual solutions enable optimal preparation for the practical realization of production. In this environment digital models are dominated. Real company model creation is and the following it's processing is not a simple matter. It requires top prepared people, corporate processes knowledge, needed software and hardware support, but also patience and purposefulness. The Digital Factory concept is the internationally renowned marketplace, with special emphasis on solutions for product development, production and integrated business processes. Hot topics are product development (PLM/CAD), production and process planning (ERP, PPC), visualization/simulation, manufacturing/automation (MES), process integration, order processing and technical sales/service (CRM). This concept can be also seen as an enterprise and information strategy managing and collaborating processes of factories in global networks. It offers methods and application solutions for product and portfolio planning, digital product development, digital manufacturing, sales and support that deliver faster time-to-value.

The digitalization approach means a general digital support of planning following the process chain from development over process and product planning to production by using virtual working techniques. It follows that the whole process of developing a new product with its associated production equipment has to be completely simulated before starting any realisation. That calls for the integration of heterogeneous processes and a reorganisation of the whole platform´s work. The thread of all process modelling and optimisation is the virtual factory life cycle support till routine operation of real production process. The digital factory approach using simulation for operative production planning and control extends the one for plant design and optimisation.

3.2 Virtual reality

Visual communiation is the most effective tool of human contact with the outside world. For analysis and decision making its necessary to generate not only pictures, which contain necessary information, but with appropriate modification achieve appropriate visibility of relevant information. Another important requirement is to see the existing context. **Virtual reality** fulfils these atributes the best. Its conceptual contribution is the 3D and 4D display and enabling a dynamic communication with the virtual world. Visual support in the virtual reality environment is much wider and more efficient than conventional systems.

The main focus in the development of virtual reality in terms of research and development of raw material should be conceptual and technical management of static and dynamic visualization of processes to solve specific problems. The first step is their handling in the laboratory conditions and following operational verification. Real appear two main

directions: using virtual reality for design support and using virtual reality for the management support. An example of virtual reality utilization is a creation of 3D models of technological equipment and their connection into the virtual manufacturing processes during the design, resp. presentations to investors. A good example of using advanced virtual reality in the management process is virtual support of managing technological aggregate, by which we can scan and identify possible errors on aggregate during the maintenance and running of the operation, etc.

3.3 Technological logistics

In terms of theory of processes and systems, we can generally devide processes into the three basic types. **Transformation** (processing and installation) represents technological operation and objective of its management is to provide planned (programmed) running of transformation. **Transmision** (transfer) and **cumulation** (storage) represent logistical operations and objective of their management are flows. From a systemic point of view, technological operations present system components and logistic operations system relations. System organization, it means that interlinking of components is objective of organization.

Technological logistics is part of logistics focused on field of technological processes. It creates the lowest level from the hierarchical point of view. Logistical processes which form a part of technological processes are running in the technological aggregate. Currenty, is technological logistics developed only as supplementary component of technological processes and does not have systemic theoretical ground. Its systemic integration into the logistics took place only recently. Logistical processes provide processes change-transformation and allow their optimal implementation. From that reason they include important innovative potential. As an example we can present logistical processes in the field of processing the granular materials. In the case of magnesite processing there are following transformation processes: processes of drying, calcination and sintering. Logistical processes are focused on the coordination and management of flows and in these field we can devide them on processes **rheological** (material flow), **hydromechanical** (medium flow) and **thermodynamical** (thermal flow).

Rheological processes characterize such movement of granular material in the compact layer. Thickness of layer represents the accumulation component and movement of the layer represents transmission component of the process. The movement can be vertical or horizontal and takes place thanks to gravitational forces, pressure forces and centrifugal forcces. Thickness and the type of movement of the layer and material in the layer decisively influence transformation processes and recently have been subject to the significant innovation. Based on this principe was designed new type of thermal aggregate working in the thin compact layer and significant contribution also was increase in the layer thickness in the rotary furnace (Dorcak&Spisak, 2004).

Hydromechanical processes ensure the movement of gaseous and liquid media through the compact layer, resp. process is performed by flux in the fluidized layer.

Thermodynamical processes include heat and material transfer as well as their accumulation. By increasing the intensity of transmission we can decrease equipment sizes, resp. increase their effectivness.

The main contribution of technological logistics is fact, that is seeking a solution to the problem directly in physical, executive field, wchich forms the essence of the process.

Solutions at the higher hierarchical levels can be optimized only based on the output from the technological area. The benefits achieved by the solutions in the domain of technological logistics can reach up to tenth percent of the process cost, which confirms that this area has great potential for the innovation. Its use mainly depends on the professional handling of the task, therefore education of the professionals and systemic research of the topic must be regarded as decisive factors.

An example of difference in the nature of technological and logistical innovation can be arrangements leading to the same change on the level of the technological process. This is achieved by decreasing a temperature of caustization with the same fuel saving for the caustization. Technological measure (change in transformation) depends on adding appropriate chemical reagent to decrease the temperature of decomposition (for example NaCl) into the burden. Negative is change of the final chemical composition of the product, which disables its use for some applications, as well as the costs of the reagent.

The same can be ensured by applying the principles of the technological logistics. Specifically, by synchronizing thickness of the layer, heigth of the zone with hydraulic and thermodynamic processes, which is expressed by longer stay of the material in the detention zone. This causes that the same degree of magnesite caustization will be achieved indeed for a longer period, but at lower temperature and significantly lower costs without affecting quality of the product and composition of the product.

In terms of using advanced technology in the logistics, new concept should be based on the PULL system, it means on the tensile principle and should meet following logistical requirements:

- eliminate the need for processes, flows, supply bins,
- integration of processes and equipment
- flow of supply bins,
- balancing capacity of resources in the supply bins,
- harmonizing production and transfer doses in the manufacturing process,
- minimalization, directness, uniformity and fluency of the material flows.

Ideal in terms of optimalization of resources would be a manufacturing process, which would work **without need of supply bins**, it means that everything would be running by system JIT (just in time – right in time). This state presents technological-organizational optimum. However to reach this kind of state in the mining company is probably not possible. Despite of that the elimination of the suplly bins need in the limited scale is real and possible to perform by harmonization of performances and capacities of the machines and equipments. And mainly through the integration of the manufacturing operations and processes into the one technological aggregate, which eliminates a need for the maintaining processes and equipments between them. As an example we can use new integrated thermal aggregate, in which we integrate supply bin, drying, dedusting of combustion gases, pre-heating, calcination, cooling, windy classification and deplacement of the product. Another option how to eliminate the need for the supply bin we can use transfer as well as mobile supply bin.

If it is not possible to remove suplly bins completely, the possibility for more effective system of extracting and processing of the raw material we can use - FIFO (first in/first out) supply bins with piston flow of material in them. In these supply bins does not occur mixing of raw materials of different quality from different doses which moving one after other. Flow supply bins work on the principle of gravity, they are simple in design and enable to

create self-organizational systems. In terms of maintenance and operation they are not so demanding. The harmonization of the stock volume in the supply bins means determining the optimal storage capacity of supply bins and estimation of the optimal level of the stock volume in them based on the needs of technological process. For this purpose i tis appropriate to use simulation and balancing models of manufacturing process. On one side resources are linked with financial issues, on the other side they are inevitable for the optimal functioning of some technological processes. They are inevitable mainly in front of narrow place, resp. in front of continuously operating aggregates, for example in front of rotary and pit furnace.

To ensure effective and smooth running of the manufacturing process, the condition is to define optimal size of production and transfer dose. Mining process is characterized by conherently-discrete material flow, conherently-discrete running manufacturing processes and till now also by variable size of manufacturing and transfer dose. Mining dose is given by extraction method and sizes of mining block, transfer dose by capacity of transferring equipment and mineral adjustment is running mostly continuously. Dose difference causes mixing of raw material different qualities. Consequently arise problems with sustainment of the productoin quality by unstable avails adjustment, capacity utilization and uneven loading of manitenance equipment. New concept of extraction and adaptation, which is based on PULL principle, requires to revaluate existing system and not only to synchronize manufacturing and transfer doses in the production, but also to adapt them to the customer requirements. The basic criteria for the optimal material flow are his **length, directness, uniformity and fluency**. Placement and organization of the process and directness influence length of material flow, uniformity and fluency is affected by level of its use. All mentioned qualities of the material flow influence its economic situation. Fulfilling the requirements for minimization, directness, uniformity and fluency during design it is possible to decrease investment costs and during the process running dut down operational costs, specificaly costs on transfer and manipulation, costs resulting from the decrease of resources volume and production in progress and also costs for maintenance etc.

3.4 Concept of Advanced Process Manipulation

Concept of **Advanced Process Manipulation** (APM) is based on procedural principle, where all handling activities are focused on optimal running of the process. This approach brings significant changes into the handling of the processes. In the present the most of solutions in some way take the process into the account., although real proces oriented approach is mostly an exception than rule. Existing individual solutions show conceptual advantages of this approach. Classical approaches can exist only in certain connection with these new conceptions. While using external manipulation, subject is dominating over an object. Object is passive and is waiting for the intervention. When using internal manipulation, object is active and requires a minimum of the external forces. Manipulation is made by transformation, mutation and adaptation. Manipulation process can be physical or logical and can run on the following manipulating levels: structura, organizational, operational and physical (Hughs&Grigg, 2008).

On the structural level is specified optimal working instrument. By structural arrangements we can approximate a process to the optimal. On the structural level is process influenced by structural components, by their interconnection and parameters. Structural component influences the process in the way that becomes its part and executes

the activity on behalf of the targeted process. It is preferable to place manipulating components on the local level, which perform their operations spontaneously on behalf of running process. Simple exponents express the laws of self-organization of irreclaimable processes. Process running based on exponential laws requires a minimum of external interventions. Real processes are generally running in some restrictions. In these cases processes are expressed by more complicated relations including simple exponents. In case of restricted resources the process can be approximated by logistical curves. However these do not have that preferable qualities as simple exponents. That is why is necessary to execute process management based on optimal trajectory.

Organizational level is characterized by choice of organizational forms, determining working mode, which insures that process is running close to the optimal mode. For specified structural level is necessary to find the optimum, or its borders. Manipulation task is to find aggregate of process trajectories, which reflect starting state into the final state. This transformation is executed based on physical laws. All processes in given device are harmonized.

In the operational level is manipulation executed by handling, which reflects real state of trajectory into the desired trajectory. I tis not possible to run the process without disturbances. Disturbances are time functions. Process shoulg be running close to the chosen trajectory. Planned trajectory is stabilized by management. Handling parameters are process parameter and product parameter. Management system or operator determine handling parameters and execute control interventions. Optimal strategy on this level is stabilization of process trajectory. Optimal working mode is reached by control devices. Priority has prediction before correction. Minimalized external handling is necessary for the execution of the radical changes of the process.

On physical level is realized macroscopic process, which determines microscopic processes. By manipulation we understand a realization of operational interventions through the active physical items based on physical laws (power, movement), which functions as converters, accumulators, etc.

Optimalization represents qualitative jump in the processes improvement. Main task of optimalization is to find from the group of extremal curves the most preferable curve. This process is connected with principal and practical difficulties. Hierarchical mathematical model of the processes here plays the key role. Optimalization of the processes in context of advanced process manipulation can be devided into the following phases:

1. Technical optimalization (decrease in restrictions):
   - designing – optimal components of the process,
   - organization – optimal ties of the process.
2. Operational optimalization:
   - planning – optimal trajectory of the process,
   - performing – optimal running of the process.

3.5 Process approach, modelling and simulation

In the process approach is key attention devoted to the process needs, which enables to be pro-active. It is necessary to create a system which is able to react on process needs, with the appropriate technology, process organization and adequate management (Davenport& Prusak, 1998).

It means to have the appropriate process elements, their interlinking, quality sensors, by which we can detect the phase of the process as well as present and future trends. In this

manner a whole system, starting from the design of the process and ending by its realization, is more and more sensitive on what occucs in the process. Process takes place on various levels. On each level process should run in the optimal mode.
Process orientation can be devided into the two parts:

- first part (analysis) includes collecting data about the process and ability to understand the process
- second part (synthesis) forms an ability to use these information to influence the process

Sensor-based approach is a philosophy, which optimalize use of the process potential. Processional orientation may present significant benefits. One of the key factors of advanced process orientation is an understanding of the process essence, which has two basic aspects:

- However difficult is the process, it is necessary to understand it.
- Big amount of sophisticated and advanced tools exist existuje, which are available to those, who have already understood them and already knows how to use them.

Modelling and simulation

An example of available sophisticated and advanced tools is modeling and simulation, which are used for better understanding of the process laws, but also for the experimentatory use with different prepared advanced solutions resulting in the process development.
Symbolic models are the abstract representation of the real world. We talk about verbal or formalized description of modelled system, for example through the graphical or mathematical presentations. They are unchangeable and uncoverable experiment tool, because it is more simple to manipulate with models than with the real objects. Models allow to understand better a reality based on following the changes of chosen indicators. Simulating model by using simulations give the possibility to experiment effects of the prepared actions during the running of the processes. Analysis of simulation results offers a good orientation about o the extent and impact of planned measurments. In the present we use two models for the purposes of better effectivness of production and process management:

- First type form **models of production systems**, which serve on the simulation of functioning and behaviour of these systems as a whole. They are usable on the strategical level for defining technical-economical efectivness and impacts of prepared investments or racionalized actions, on tactical-operational level for planning and shcedulling of the production etc. An example of this type of model is blancing model of the manufacturing process.
- Second type of model represent **models of processes and devices**, which are used for solving certain types of cases, for example for design of thermal aggregates. As an example of this group serves model of thermal adjustment of magnesite in the rotary furnace.

By connecting both above defined model types we get an advanced hierarchical model system, which is based on hierarchical principles and allows directly study complicated processes and on this principle perform also their optimalization. Hierarchical nature of the processes is generally known and accepted. To each hierarchical leved corresponds its specific process, which includes elementary processes and their combinations. To every hierarchical level applies specific type of movement. Technological process are expressed

by evolution degree in time. Process on certain level we may consider for an abstraction of complex process executed on various levels. Processes on the individual levels are reciprocally interconnected in that way that processes on the lower level perform processes on the higher level. This is reflected in the reciprocally corresponding parameters. Kinematic parameters on the lower levels correspond to the force functions on the higher levels. Processes on the higher levels determine processes on the lower levels (Prawel, 2007).
On each hierarchical level are internal ties and external ties between processes on the individual levels. Those can have different structure and different complexity. They can be hard or soft. Soft ties express an autnomy of the processes. Ties between lower and higher levels are securing and between higher and lower levels are controlling. Controlling and securing ties are also between the components on individual levels. Process innovation can take place on intro-level and between-level. Typical hierarchical levels of the processes are: **designing level, managing level and physical level**. On physical level are running material processes based on their own laws. On the management level is running controlling process, which provides process on physical level by determining operative parameters. On the design level is proposed executive (material) and managing process. Designing process has only free ties towards the proceses on the physical and operative level. Between physical and managing level are strong inter-level ties.

3.6 The results in the research and development of technologies used for the the raw material extraction and treatment

Dealing with the research and development of technologies used for the the raw material extraction and treatment is the content of ETP SMR Strategic Research Agenda (SRA) which shows the way the mineral industry should proceed in forthcoming decades if it is to serve European society in the way necessary. The structure is divided into the the 4 focus areas. The scope of the established focus areas relates directly to particular steps in the raw material value chain. The focus areas covers the whole life-time of a particular product, from exploration and extraction until reuse and recycling. It reaches processes from the exploration, the identification of valuable mineral resources to the sellable products. All steps of the supply and production chain for mineral resources are underlined with societal issues of various kinds.

- **Exploration, Extraction and Closure (**Towards Total Resource Utilisation, Energy efficient fragmentation technologies, Innovation for materials handling and logistics optimisation, Internal processing systems for re-use and recycle**,** Environmental footprint reduction using new processing systems, techniques (life cycle assessment)**,** Knowledge building networks)
- **Reuse & Recycling (**Information network for mineral and metallurgical industry, Industrial network on waste prevention and recycling aiming at turning wastes into products, Prevention of waste by innovative processing - innovative processes turning waste into products, Feedstock recycling (plastic, waste wood, chemicals), Footprint free production - Recycling of materials and better use of mineral resources
- **Products & Materials (**Creating new mineral product functionality through an enhanced product and customer understanding and knowledge building**,** Finding new application areas for mineral products and designing the mineral products for tomorrow, More efficient management of innovation in the mineral industry and building new products development capabilities)

- **Mineral Economics and Societal Issues (**has identified research areas in close relation with other focus areas to detect and use cross-sectoral synergies.)

3.7 Customer oriented production – pull system of material flows

For the management of the manufacturing process in general exist only two philosophies. Principle of the first pressure philosophy is to push the fastest and most effectively all resources, material, semi-product through the whole manufacturing chain and through that gain the most of the product not regardless following business activities, mainly consumption (production on stock). Sometimes we call this procedure the push method. Simply we can say, that type of systems „Push" are those manufacturing systems, in which is production managed with a stiff plan (production on stock). Here the priority play business objectives (for example maximalization of using company resources, minimalization of company costs etc). Methods of push management can be various, they always depend on centralized monitoring and influencing of individual activities. This method is characteristical for the first-fabricated industried, it means industries with the uniform manufacturing process. Typical example of these processes are processes of exploitation and processing of the raw material.

Second method is puulling, based on the opposite principle. Using this method - **„pull" method** impulse for the production-logistic chain comes from the customer (custom made). In the moment, when the final process is requested by customer for the delivery of the product, he will turn to the previous process with the request for delivery of the necessary inputs, this process requests previous, etc. This method i sused as an advantage not only in the production and delivery of the cars, but also computers and other goods, which are configured based on customer requirements. For those are parts, semi-products mainly „pulled" by flow, than pushed in the fron in big amounts according to planned directives. Management of the material flow gets more simplified.

Both philosophies have their advantages and disadvantages. Manufacturing processes and conditions of their functioning predict form of a management and by that also choice for one of these philosophies. Till now in the mining processes was typical to use PUSH system, which corresponded to the character of the manufacturing process and conditions of its functioning. The problem is, that conditions have currently and significantly changed, customer is not waiting passively for the delivery of the ordered product, suppliers of the input into the mining company are also following market rules. Meanwhile also the technologies used in the mining industry are changing lately and are adapting to the energic, technological, ecological and social requirements. By that way they create conditions for the implementation of more economical PULL system of the management also in the mining industry. While most industries have undergone, resp. proceeds on the more effective energic, material, personal (PULL) system of operating manufacturing processes, mining industry (partly with a reason – natural conditions, technology of extraction and adaptation etc) is using only classic (PUSH) system. If the concept *extract - adapt - sell – to dispatch* changes on *sell- extract- adapt- to dispatch* It will make signigicant changes in the companies management, which would bring new possibilities of increase business competitivness, targeted exploitation and targeted processing of extracted material and it will significantly improve customere service.

Application of PULL system in the planning and managing of the mining activies opens new possibilities in optimalization and better effectivness of the business processes.

Utilization of this new system in the original structure of business activities is not possible, resp. it would enforce significant corrections, which would seriously restricted its innovative potential. For this reason it is inevitable before application of new system of management execute radical (reengineering) changes in organization of existing corporate logistic systems of mining companies.

All known mining production processes as well as the majority of homogeneous manufacturing processes is organized by PUSH system, it means based on manufacturing process and mainly its narrow place and is not based on customer needs, which are in this system on the second place. Main features of this type of organization of the production process are:

- Maximalization of production volume by PUSH system of management of manufacturing process (output restricted only by transmittance of narrow place),
- Minimalizaton of unit product cost by decreasing fraction of fixed component,
- High costs for storage and manipulation given by creation of between-operational resources,
- Advantages connected with the forcus on one, resp. narrow selection of goods,
- Possibility of flexible and immediate product withdrawal from the shipping stock by customer,
- There must be fullfilled the requirement for tzv. limitless withdrawal (resp. stock), it means that is necessary company´s ability to sell all what was produced. This is possible only with a permanent dominance of demand over the offer,
- Problematic enlargment of production range,
- Simple and stereotype system of producton planning and managing,
- Feedback of quality control of manufactured production,
- High degree of using labour force, high labor productivity

PULL system is based on the opposite philosophy. Needs of customer are dominant and manufacturing process must be organized in that way, that i tis able to fulfil customer needs in flexible and meanwhile effective manner. Basic characteristics of this system are:

1. maximalization in fulfilling customer requirement – hard focus on the customer,
2. ensuring production flexibility and ability to react the fastest way on changing requirements of the customers (changes in production selection and amount of whitrawed production),
3. Pull work organization, by which are parts „pulled“ by manufacturing process based on need to finalize customer orders,
4. Management of fluency of material flow and its synchronization allow to have quite better overview opposite to classical PUSH system about material flows and stocks,
5. Minimalization of running times of production by choosing the extent of production doses in that way, that there will be minimal between-operational times, mainly times for preseting the machines and downtimes,
6. In building of quality control into the manufacturing process and its consistent exercitation, quality is not controlled because is produced,
7. Minimalization of between-operational stocks by synchronization of produced doses in accordance with requirements of purchasers will allow cost decrease for storage,
8. Decreasing variable costs caused by decrease of measurment of production in-process,
9. Increasing fixed costs because of not even loaded process capacity,
10. Decreasing of oversized work pace using creative potential of human labor and team work,

11. Information support of processes and transparent information flow.

Based on comparing basic features of PUSH and PULL system is possible to say, that in regard to changed market conditions, application of PULL system in the specific conditions of mining industry can bring not only significant decrease in total production expenses, but it can be the way to ensure its permanent maintainable prosperity.

3.8 Ecology and safety

Dealing with the research and development of technologies used for the the raw material extraction and treatment is the content of many European project ideas using advanced technologies to reach the objective are listed following:

- New technologies for management of gases which are rich in SO2 aiming to produce by-products, e.g. gypsum directly from gases
- Environmental footprint reductions by developing new technologies and applications: water treatment, gas streams handling, etc.
- Radical changes and innovations in mineral and metallurgical processes to improve efficiency and decrease environmental negative impacts
- Clean processes (hydro, bio, pyro) for treatment of complex ores and wastes aiming to reduce environmental impact
- Materials and chemicals to reduce environmental footprint
- Monitoring tools and sustainability environmental management standards, indicators
- Innovative methods for disposal of tailings
- Technological and administrative tools for reduction of mining waste
- Management and disposal of wastes in mining operations
- Assessment of the environmental impact of mining activities on groundwater and soils - evaluation of the current knowledge base, mitigation or remediation technologies
- Rehabilitation and chemical and biochemical processes for extraction, sequestering or stabilisation of pollutants from contaminated land
- New technologies preparing shafts for filling; old shafts, excavations, shallow deep mining activities
- Remote sensing technologies for assessment and monitoring reclamation process
- Pro-active policies of the mineral industry implantation in developing countries to improve and sustain regional development

4. The 21st century mining corporation concept

"The 21st Century mining corporation" concept presents a complex objective created and target-oriented solution for mining industry, making provision for all relevant innovative megatrends, integrating latest results of research and development in the form of progressive technology, system processes into **the integrated holistic solution** flexibly adaptable on the conditions of factual mining corporation.

This integrated holistic solution (Fig. 2) incorporates the full raw material value chain - raw material sources searching, their extraction, primary and secondary processing, up to the product finalization. It makes provision for the product re-use and recycling on basis of minerals, territory revitalization after mining activities, but also an information support of mining enterprising, system control and mining corporation logistics, including all necessary utility and corporate processes (maintenance, transport, etc.) and socio-economic aspects.

The solution rises from the fundamental dividing of extraction and raw material processing value chain into the five basic research areas according to European Technology Platform for the raw material extraction and treatment. that details, fills up and target orientates goals for filling the concept vision of intelligent, complex and effective raw material extraction and treatment by an advanced technology into the new products with higher added value, all allocated underground – that is „Invisible Mine" vision – an vision of not only underground extraction, but also underground mineral processing. This vision presents the mining and processing industry plants, that can be "invisible" not only from the reason of their underground location, but mainly from the point of absence of their unhealthy impact against environment, miniaturization and smartization, minimizing consumption of energy, an advanced system control based on the principles of APM (process self-steering, self-regulation and self-organization), complex non-waste extracted raw material processing into the products according to customer requirements. To create such a solution it is needed to obtain the critical amount of information and knowledge.

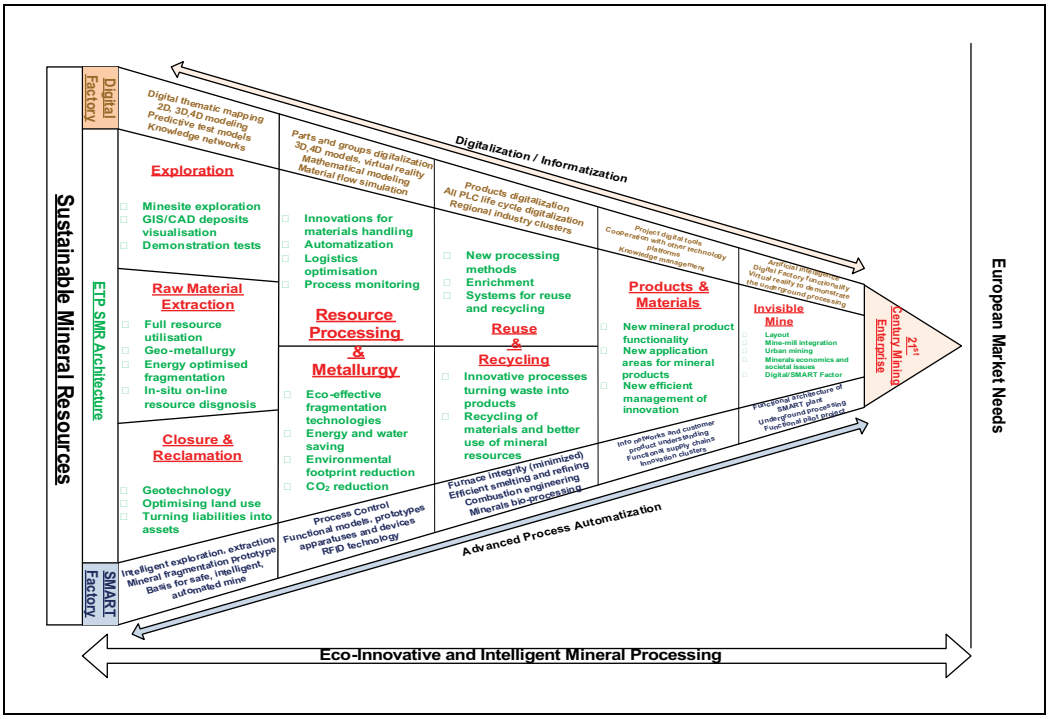


Fig. 2. The scope of the 21st Century mining corporation concept

The task to use advanced technologies in mining industry links together more economical and effective utilization of accessible raw materials with emphasis on the environmental safekeeping at the same time. This complex problem demands to optimally realize wide scale of activities, starting with geological exploration, extraction, raw material treatment – up to the product finalization in two reciprocally influencing levels – in real level (physically) and in virtual level (digital). From that reason the 21st Century mining corporation concept is within digital world represented by Digital factory concept and within a physical appearance by the SMART Factory concept. The diagram (Fig.3) presents the mutual interconnection of these two worlds.

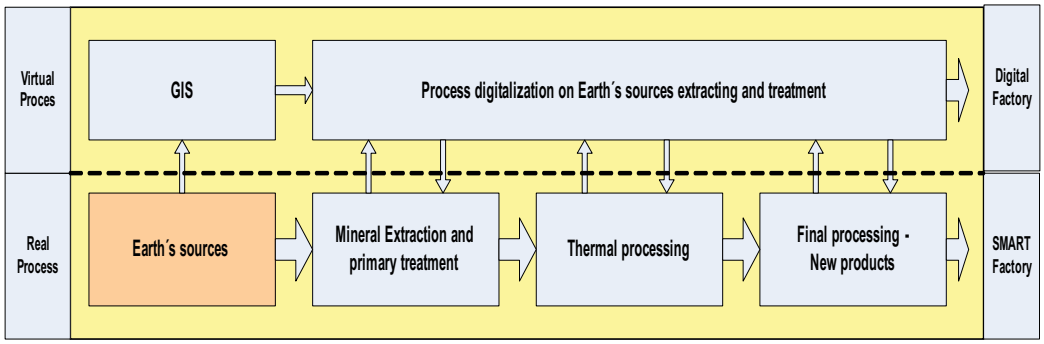


Fig. 3. The basic diagram of the 21st Century mining corporation

4.1 The Digital Factory

The Digital Factory (DF) is an advanced technology of a real production virtual picture, which shows the production process at virtual environs. The Digital factory concept serves first of all to plan, simulation, management and production optimization. Informatization, digitalization and virtual reality create the fundamental prerequisites for the digital factory creation (Kuehn, 2008).

To create the complete digital factory in the area of raw material extraction and treatment it is needed to digitize processes gradually, which digitalization would enable to reach important benefits and these can be gradually enlarged and integrated. From this point of view the digitalization of two corporate processes is actual: specifically corporate processes aggregating scheduling within logistics and key technological aggregates (e.g. thermal aggregates) within technology. These processes present a suitable basis for the progressive digitalization of others referring processes. The key technological (e.g. thermal) processes digitalization within the raw material modification enable to utilize a predictive approach in their control and based on the mathematical simulation models it enables to use a virtual techniques in their control. These processes represent from technological point of view bottlenecks, in most cases. The modelling and consequently also operational harmonization and integration is performed by progressive foregoing and consecutive process models linking into the thermal process models.

The final digitalization objective is to create **the whole technological process model** vertically integrated with **superior advanced corporate planning and control system**, which integrates into the one system all corporate processes on the tactic-operative level. Thereby we can reach the corporate process integration – the foundation-stone of Digital Factory concept.

The advanced corporate planning and control system is based on the deposit Geographic Information System utilization, the Hierarchical mathematic-simulation model of the raw material extraction and treatment process and the aggregate planning system, which creates the superstructure above the former two models. Such a system makes it possible a mining operation functioning in new effective mode, fully respecting customer needs, the possibilities of raw material treatment and processing, as well as the extraction possibilities and constraints. On the following diagram (FIG.4.) it is displayed the interconnection among an individual parts of the proposed system.

Each from the sub-models is able to work independently, what may bring a considerable contribution, but only en bloc it can reach and fetch all consequential synergistic effects. The scope of the Aggregate Planning System (APS) is to process customer-tailored requirements

and creates a suitable custom-made fulfilment for production process regarding its constraints. Through the right harmonization of all activities, the APS can provide production processes from material input to their outputs - products. Through the APs interconnection to the Hierarchical Model of Production Process (HMVP) and Geographic Information System (GIS) by reciprocal information flow the system can ensure more perfect and effective control for full mining production processes.

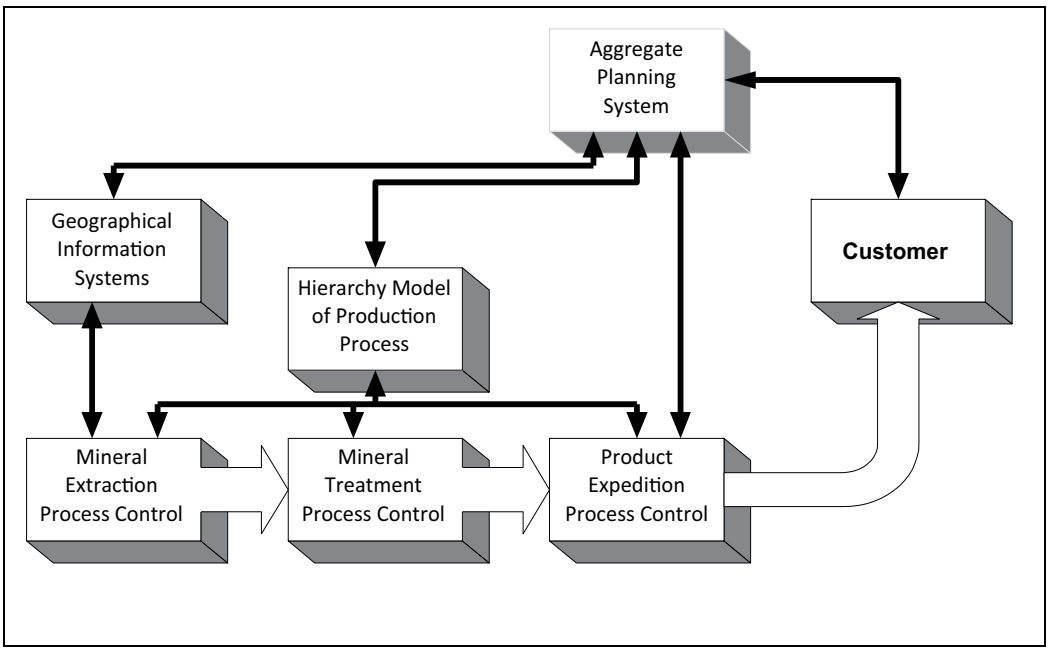


Fig. 4. Advanced corporate planning and control system

4.2 The SMART factory

The target of so-called **SMARTization** is to create an intelligent and sophisticated factory in area of the raw material extraction and treatment and to ensure its functions within full life-cycle, which comprise a designing support, planning of operation-service activities. Physically version concept realization of „invisible and intelligent" factory - mine will be realized in order to fulfil requirement to be up to standard of **4e**: must be effective, economic, energetic unassuming and ecologic.

Smartization decreases constraints lying on a process and ensures their realization by natural style with the outer control intervention minimization and full elimination. Performed processes run over the surrounding of their technological optimum. At the same time it is possible to simplify technology, equipment and process organization. Hierarchical approach makes it possible to perform process optimization independently of structural, organizational, operating and physical levels. The structural level task is to decrease constraints on process in order to run over maximally freely with the maximum internal feedbacks, which meet ideal technology. At the most processes it is possible sort of to come near to this stage, which we can denoted as technologically optimal process. Technologically optimal process is able to be a measure for the individual alternatives. The optimization goal is not only to find optimal

output of the process, but also optimal course of the process. On the organizational level we can designate an optimal trajectory of the process for specified conditions of the process within the frame of its possibilities. Operational level ensures the course of the process nearby an optimal trajectory. Under the optimization from the point of the scope we understand not only finding of the total optimum, but also relative process improvement.

Application of SMARTization to the raw material extraction and treatment processes will designates a meaningful upgrade of technology. We reach benefits in operating costs decreasing, mostly in energy savings and in the higher raw material assessing. It will be created a possibility to develop new products this way or to increase the add-on value of existing ones. The increased contribution is estimating on 25-50%. The master impact will be on the potentialities of transition from standard production into high-tech category.

SMART factory concept is created gradually, based on the advanced technology ensuring the real semi-operational and operating activity. The realization of complex SMART technology implementation within the frame of full corporate processes spectrum needs to create partial solutions alternatively innovated technological islands and integrates then consequently into the total and functional solutions. So-called SMARTization will comprise the modifications of real and virtual objects (mathematical models), database and visual communication on the level of a partial operations and components through the machines and aggregates up to a complex operating technology.

1. **Partial operations and component parts** – development of intelligent partial operation – equipment component part, aggregate or technology with elements of sophisticated autonomous operations, e.g. autogenous grinder component.
2. **Equipments** – represent development of autonomous equipments built from Smart type of component.
3. **Aggregates** – represent development of bigger unit as are individual equipments and dealing with integration them into the aggregates, e.g. windy separation integration into the swing-hammer crusher.
4. **Technology** – represents development of the biggest unit, components and aggregates integration into the sophisticated complex, which represent the technology for factual technological process or nodal point, e.g. technology of thermal processing in thin layer. In this manner the created advanced technology will designate the transition from automated equipment supported by control system to high intelligent autonomous self-steering system equipments without assistance of control system! The final stage is to establish the integrated and smartized technology complex of raw material processing.

The strategic goal of smartization is to reach sustainable development in area of raw material treatment. It can be reached by securing the database and knowledge for all life-cycle phases by the process and aggregate research and development, with the aim of to minimize dimension, technical efficiency, by the technological logistics (TL) and the Advanced Process Manipulation APM principles applying. The solution is based on the research and development of advanced technology within raw material extraction and treatment within the approach of Invisible/Intelligent Mine and on virtual designed functional verification of advanced technology for the raw material extraction and treatment process application.

The application of up-to-date science research results in the form of production technology and logistics area SMARTization demands to apply adequate approaches to control corporate processes in the concept design. The systemic and process approach will cover all four main production process characteristics: *quality, quantity, time and position*. Production process on all levels will be divided into three main groups namely:

- transformation – processing, which causes the quality, quantity and time change,
- transfer - transport, this causes the position and time changes,
- cumulation - storage, which causes only a time

At present time the dominant control system is **the combined system** in production processes, built-up from program, forward and feedback control (Lipsett et al., 1998).
While in technological processes dominate programming-feed-back control, in logistics processes dominate programming-forward control. The progressive competitive tension impact and requirements on quality production, as well as on system flexibility we need to put the accent on predicting component within the control. This is one of the reasons to prefer logistics and process approach within the control on the present. Upon this trend also the managerial structure proposal for the new concept of the raw material extraction and treatment process with using of advanced technology is responding, which prefer the programming-predicting control, an advanced manipulation process concept (self-organization, self-regulation, self-steering), the control informatization a digitalization and the logistical principle into lowest hierarchical process levels implementation (technological logistics).

5. Methodology for the accelarated transfer of advanced technologies into the practice

In order to accelerate a transfer of advanced technologies and processes into the practice was designed original organization approach towards research and development activities at our working place. It is built on the expansion of research space towards implementing activities and replaces classic activities before realization (project preparation, test operation, start up of the operation) by advanced tools (virtual reality) allowing greatly speed up the innovation process and ultimately reduces investment and operational costs of modern technology. This methodology is schematically shown in the picture below Fig. 5.
Proposed methodology is designed in regard to the up to now results of realized applied research, resulting in the creation and development of advanced technological processes and technology in the science field of exploitation and processing of raw material applied in

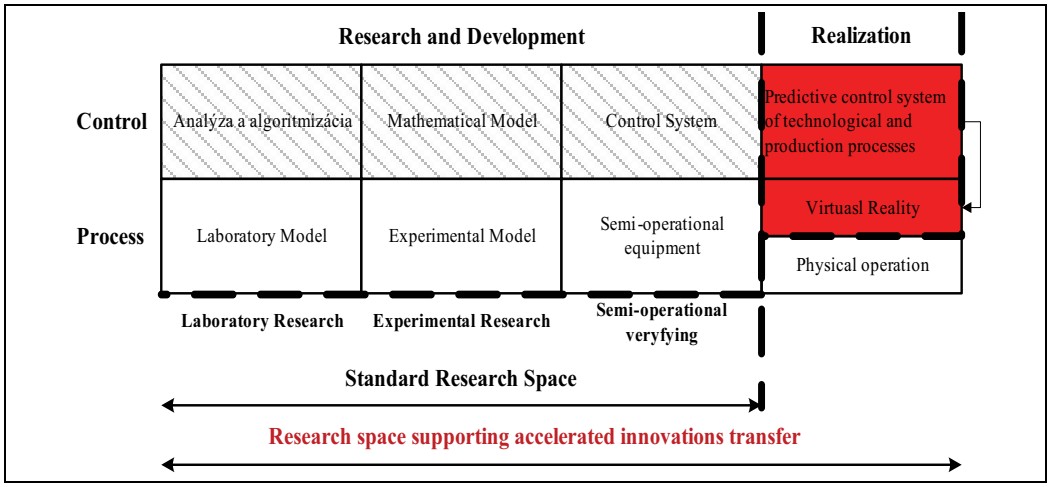


Fig. 5. Concept of extending the research area which supports accelerated transfer of the technologies.

new **extraction and complex magnesite processing technology** pilot project . Purpose of this concept is integration and finalization of up to know partially solved research tasks focused on innovation of processes and tools for acquisition and processing of raw materials to the functional and half-operational verified scheme of complex high-tech technology processing of magnesite.

In the context of the above mentioned approach towards applied research linked with a transfer of new knowledge into the practice is particularity of draft concept for **complex semi operational verifying of new technology functionality without the need of its physical construction**, which brings very significant cost and time savings. Concept will be handled locally in a semi-operational environment also the functionality of particular technological processes and their compatibility with partial mathematical models of processes and aggregates. Consequently on complex mathematical model created in virtual reality environment will be verified whole production process together with created intelligent monitoring system, predictive control system of technological and manufacturing process and the proposed business logistics system.

The advantage of this methodology compared to classical concept of access to research and development is:

- half operational verification of individual technological processes (thermal and finishing) in physical form, at local scale and not depending on time
- verification of the proposed logistics processes in the operating conditions only in virtual reality
- predictive control system of individual technological processes integrated in hierarchical management system of manufacturing process will be verified in operating scale in virtual reality
- verified links and optimal allocation of technology and logistics operations in the new technology prior to its construction
- defined the precise techno-economical parameters of future operation

6. Conclusion

The target of this contribution has been to analyze complexly the potention of advanced technology application and innovative trends for making the raw material extraction and treatment process more effectively and to ensure the competitiveness and sustainable growth of mining corporations. The solution of presented goal suggests the innovative solution application in specific and exacting conditions of mining production processes, the complex appraisal of their contributions and the generalization of gain experience from its application. From the aspect of innovation character were appraised technological, logistical and economic-organizational innovations.

The results from an analysed impacts of suggested innovation arrangements, as well as the present experience from its applications in practice were utilized for a model method definition to support innovative process, starting-point and principles on innovation arrangements proposal with the highest level of changes, to the conceptual model proposal of new advanced technology on raw material extraction and treatment area – **the 21st Century mining corporation concept**. This concept is the generalization of developed and progressively applied concept of a complex magnesite ore processing that contains the proposal of new productivity-technological, logistic and organizational–control system. The proposal of production-technological system come out from new technology of raw material thermal processing possibilities (ITA technology, micro-fluidic high-speed rotary furnace) as

well as new extraction and finishing solutions. The logistics system proposal is based on an application of PULL system for controlling corporate processes, possibilities of address extraction and processing according to customer-tailored requirements, which come out from aggregate system planning concept, possibilities to utilize mathematical modelling and technological logistics. At the proposal of organizational–control system were utilized the forward control principle in maximum measure, process informatization and digitalization and virtual reality. Cross-sectional approach was used at new raw material extraction and treatment proposal and the philosophy of advanced process manipulation.

Submitted concept of the advanced technology application on the raw material extraction and treatment area comes out from our experience of factual own innovation applications on various level and various areas of corporate processes. On levels from process parameters change, through their restructuring, radical reengineering changes pending the proposal of new complex technology. Most listed examples were already realized in practice, or in the near future time the realization is prepared. The application of suggested solutions and concepts in corporate practice will enable corporations the greater orientation on customers, permit an effective exploitation of mineral resources, or enables on qualitatively higher level to operate with corporate sources. Significant solution benefit represents also possibility to allocate a fact, that concept results after verification in practice would be immediately integrated into educational process. Implementation of a verified knowledge about raw material extraction and treatment innovation process contributes towards an enhancement of educational-training activities predominantly at close phase gradual study a post-gradual study. The student preparation like specialist – professionals for innovation enterprising is decisive for securing sustainable development of corporation functioning in raw material extraction and treatment area.

7. References

Davenport, T.H.; Prusak, L. (1998): Working Knowledge, *Harvard Business School Press, 1998.*

Dorcak, D., SPISAK, J. (2004): The reengineering methodology utilization within a complex raw material extraction and treatment process optimisation, Acta Montanistica, ISSN 1335-1778

ETP-SMR SRA (2009), The Supplement to the Strategic Research Agenda of the European Technology Platform for Sustainable Mineral Resources (ETP SMR), *ETP SMR Brochure.*

ETP-SMR (2005), Vision Paper for a European Technology Platform on Sustainable Mineral Resources (SMR), *ETP SMR Brochure*

Hughs, T.R., Grigg, N.J. (2008): An underground Processing Plant for narrow Vein Mining, *Ballarat, Vic 2008.*

Kuehn, W. (2008). Digital Factory – Integration of simulation enhancing the product and production process towards operative control and optimisation, *White Paper,* Wuppertal.

Lipsett, M.G.; Ballantyne, W.J.,Greenspan (1998), M.: Virtual Environments for Surface Mining Operations, *CIM Bulletin*

Prawel, D. (2007). The Advent of Visual Manufacturing, *White Paper*, President & Principal Consultant, Longview Advisors Inc.

Särkkä P., (2008). The Intelligent Mine, *Proceedings of congress,* the concept of Helsinki University of Technology, Mining Congress, Katowice

15

Augmented Reality System for Generating Operation Program of Automobile Assembly System

Hong-Seok Park, Jin-Woo Park and Hung-Won Choi
University of Ulsan
South-Korea

1. Introduction

In recent years, the computer graphic technology has been growing dramatically with computer industry growth. In this exigent situation, the research of VR (Virtual Reality) and AR (Augmented Reality) technology which describe realistic scenes is performing actively (Daniel et al., 2007; Stephen et al., 2008; Bimber et al., 2008; Ong et al., 2003). Moreover, not only computer science companies but also many manufacturing companies have been studying about VR technology, because the global market requires a variety of product and shorter life cycle to fulfill the diverse demands of customers to survive in the turbulent and competitive market.

In the case of VR based digital manufacturing technologies, it can be used to analyze static and dynamic system behavior at all stages of manufacturing system configuration as an advantage (Günter et al., 2005; Park et al., 2008; Kim et al., 2008; Park et al., 2007; Park et al., 2006). To put it the other way around, the modeling work requires much expenses and efforts, because the whole system has to be modeled. Some manufacturing companies which realize the weaknesses of VR based digital manufacturing technologies have been studying AR technology.

AR means that the execution or planning ability of a real world is increased by the superimposition of virtual object on it. Basically, Augmented Reality is configured with three requisites as a real environment, markers and virtual objects (Fig. 1).

The ambitious goal of AR is to create the sensation that virtual objects are present in the real world. To achieve the effect, software combines VR elements with the real world. Obviously, AR is most effective when virtual elements are added in real time. Because of this, AR commonly involves augmenting 2D or 3D objects to a real-time digital video image (Stephen et al., 2008). AR technology can remarkably reduce the modeling work, because it uses the real environment to design and to plan manufacturing systems (Lee et al., 2008; Park et al., 2008).

AR technology is consisted basically of three modules which are image processing, tracking and rendering. The image processing module has to support the various types of the image data and can process the lots of image data. The image processing module sends the obtained image data to the tracking module. And the rendering module performs image

superimposition. Among these three modules, tracking and registration problem is one of the most fundamental challenges in AR research. The precise, fast, and robust tracking of the observer, as well as the real and virtual objects within the environment, is critical for convincing AR applications. Because of this reason, much research effort is spent to improve performance, precision and robustness of tracking. Besides tracking, real-time rendering is also basic element for augmented reality. Since AR mainly concentrates on superimposing the real environment with graphical elements, fast and realistic rendering methods play an important role (Bimber et al., 2008).

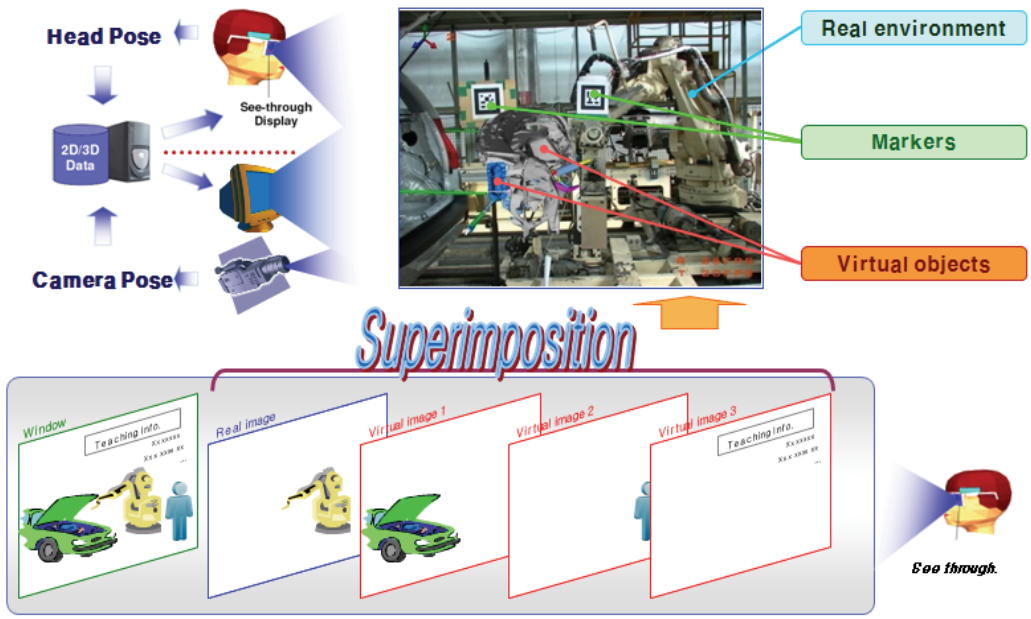


Fig. 1. Configuration of Augmented Reality System

To develop an AR system for manufacturing application, the design specification is derived out from the analysis of manufacturing problem and application environment. Based on the specification, the system architecture is designed. The necessary function modules are developed from the architecture. For the convenience of users, the application modules are also developed. To implement an AR system, the developed modules are integrated.

Through the integration of the previous developed modules, the AR-system was implemented. These user friendly UI and the functions support a planner or an operator to carry out the results easily. For proving the functionality of the developed AR-system, it is applied to a practical problem, i.e. placing spare tire on a mount hole in trunk. For this purpose, A test bed with 1/5 scale of a real system was realized. For solving the inserting problem with the AR-system and for carrying out the inserting process, the real system components and the virtual objects were developed and modeled. With the previous mentioned installation, a programmer generate the robot program for the inserting process with teach pendent in front of a monitor, i.e. the programmer carry out the programming process without seeing the real system.

2. Design of augmented reality software for manufacturing system

2.1 Design specification of the AR-System

The software development activities of AR for the manufacturing system consist of requirements analysis, specification, software design, implementation, testing and maintenance. These processes were performed recursively with correction of each step. In the requirements analysis step, the general requirements were gained from the client and the analysis of the scope of the development was determined clearly. The analysis document of requirements includes the flow chart of overall tasks, the systematical analysis information of functions, activities and data.

The concise software development scope is as follow.

1. The user selects the camera devices which are connected to computer by the device enumeration dialog box.
2. The user inputs the each number and total count of markers.
3. The 3D objects (.wrl) are loaded to use for the parts of the manufacturing system.
4. After markers and virtual objects setting, the display device shows superimposition of real-time images.
5. The translation, rotation and scale data of each virtual object can be controlled by the user.
6. The clipping plane for collision detection can be created and also can be controlled by the user.
7. The visible and invisible functions for the images and virtual objects are able to operate.
8. The marker tracking start and stop functions can be used during the software activity.
9. The user can confirm the information about program activation by the message boxes.
10. The information of each marker and each virtual object data can be removed by the user.

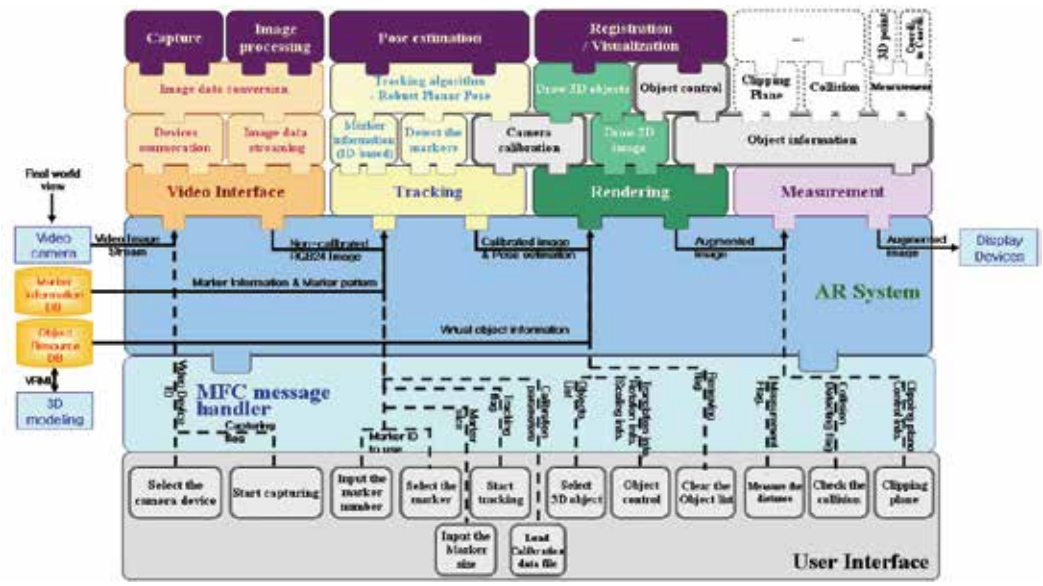


Fig. 2. Architecture of AR System

Based on the specification document of requirements analysis and the three basic modules which were introduced at chapter one, the auxiliary functions can be supplemented for

manufacturing environments. In order to overcome dim environment condition and the weakness of the virtual object, the convenient functions which are clipping planes for collision detection and threshold control panel for the dim condition. And these functions support the user to work efficiently. Fig. 2 illustrates the architecture of AR system. And the core development tool is MFC (Microsoft Function Class) based on C++.

The image data from a camera device is processed in video interface for the marker tracking. The video interface converts the video image stream into non-calibrated BGR24 image. Then the tracking function calculates continuously coordinate data of the markers to detect location of each marker. The basic coordinate system for positioning virtual objects is established through the matching procedure between the information of the marker database and the input data of the user interface. The rendering function performs generating and removing of virtual objects with calculated coordinate data. Also, the coordinate transform of virtual objects is executed by using three translations data, three rotations data and three scales data.

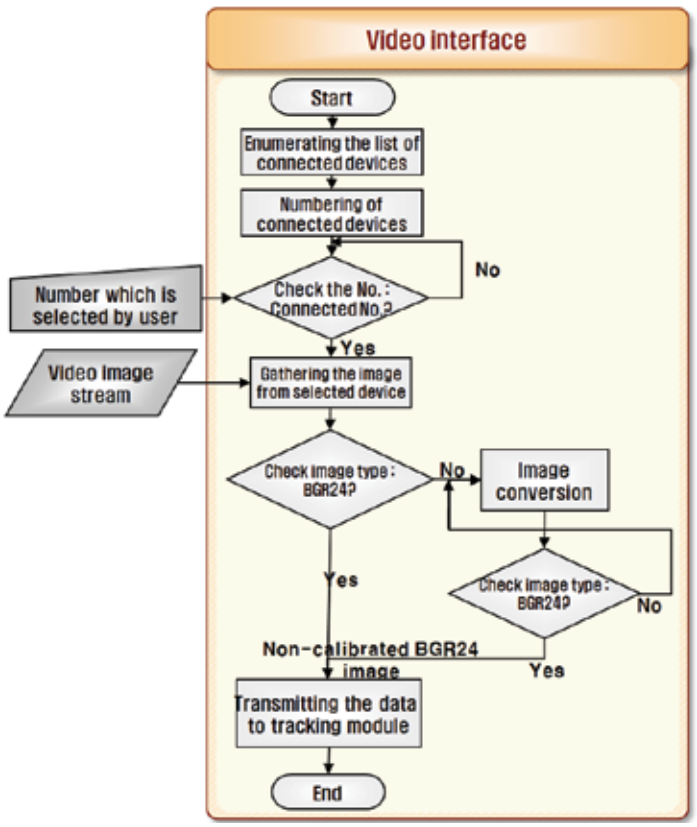


Fig. 3. Gathering the image data from connected devices

The Video Interface module executes to obtain images from external devices such as video camera etc. If the cameras are connected to the computer from the outside such as USB web-cam, IEEE 1394 camcorder etc., Video Interface module enumerates the connected devices. After numbering work for each camera, the Video Interface module compare the number of

connected devices with the number of selected device that a operator will use. If the numbers are exactly matched, the Video Interface module gathers the image from selected device to use for superimposing (Fig. 3).

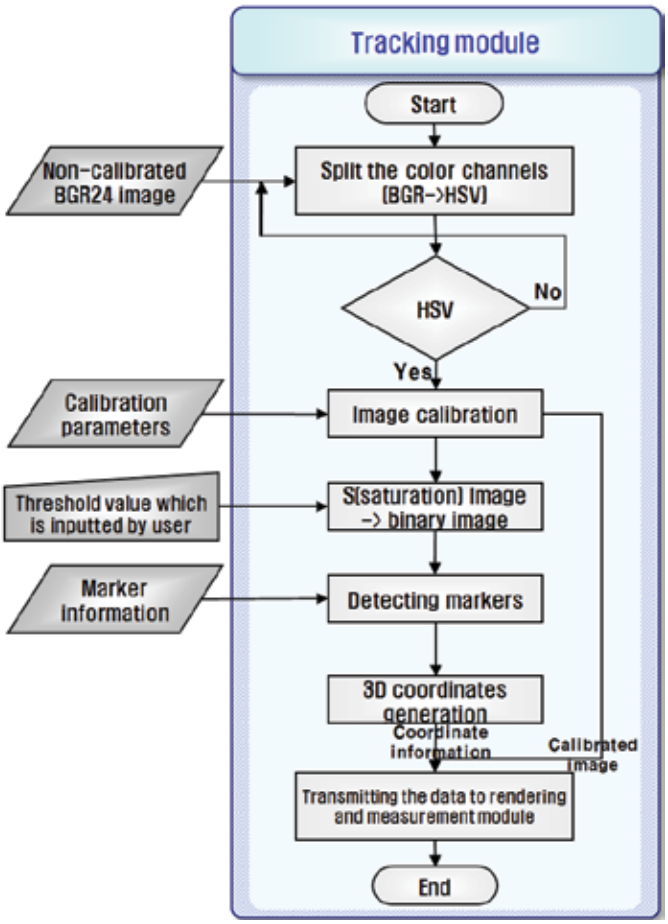


Fig. 4. 3D coordinates generation after marker detection

For establishment of coordinate system, the tracking module is used. In order to calibrate the BGR24 image that is from the Video Interface module, The Tracking module splits the color channels to HSV type. After that, the HSV image will be calibrated. The calibrated image will be changed to the binary image in order to generate 3D coordinate information (Fig. 4).

The Rendering module is to carry out the calculation process for displaying images. This module generate the world coordinate system and set the camera view position. In order to superimpose the 2D image and the 3D objects local coordinate based on marker posion and world coordinate will be synchronized. And each image such as 2D image and 3D objects will be drew on the same display plane. In case of the binary image, it will be used to change binary value to overcome dim condition of manufacturing area (Fig. 5).

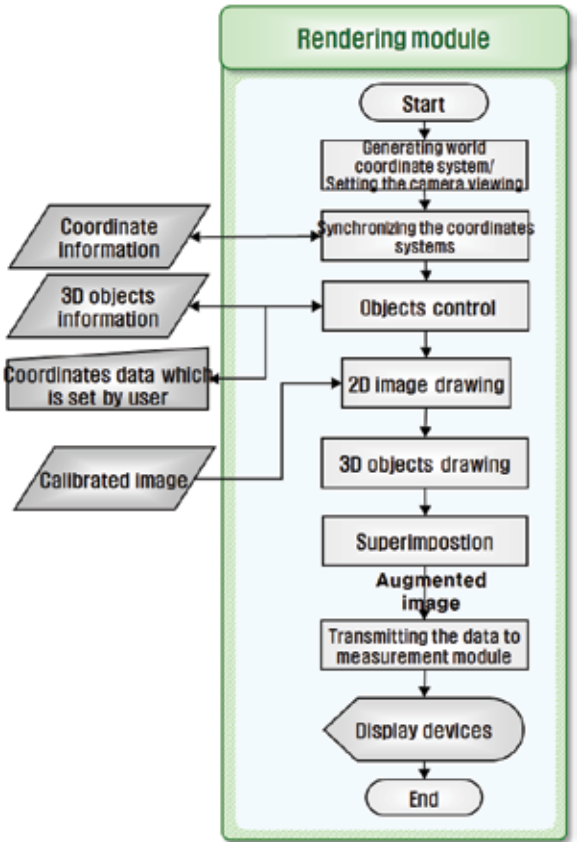


Fig. 5. Superimposing the 3D objects on the 2D image

2.2 The configuration of the user interface

The user interface designed based on basic functional analysis of whole system as shown in Fig. 6.

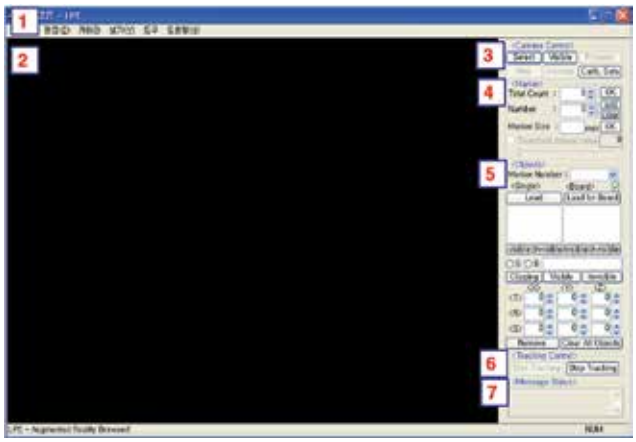


Fig. 6. Layout of user interface

1. The menu bar
2. The display plane for the render scene
3. The camera device controlling panel
4. The marker information setting panel
5. The virtual object controlling panel
6. The tracking function panel (start/stop)
7. The message enumerating panel to check program state

The menu bar consists of the functions which are performed from the right side panels. The display plane is the drawing board for superimposition of scenes and it draws 3D objects and 2D images from the camera. The user can confirm the list of connected camera devices by using the camera control panel and can change the camera properties easily. The marker information setting panel is used to input the information of the marker such as the size of the marker, the number of the marker, the total count of the markers to use and so on. And it is also used to control the marker information. The virtual object controlling panel is used to load the 3D objects for rendering and to control. The tracking function panel is used to control start and stop for the position tracking function. The message enumerating panel is helpful for user's activities to check program state.

3. Implementation of augmented reality system

3.1 Obtainment and conversion of real-time image data

Precision and accuracy of target tracking depend on the image resolution of the camera device. USB Web Camera which has low resolution is used generally for experiment at a large number of laboratories. However, ordinary manufacturing system requires the tasks with high precision and accuracy. Because of this reason, the digital video camcorder that is connected to computer through the IEEE 1394 interface was used. OpenCV is the prevalent open source vision library suitable for computer vision applications. Fig. 7 depicts whole image processing step structure.

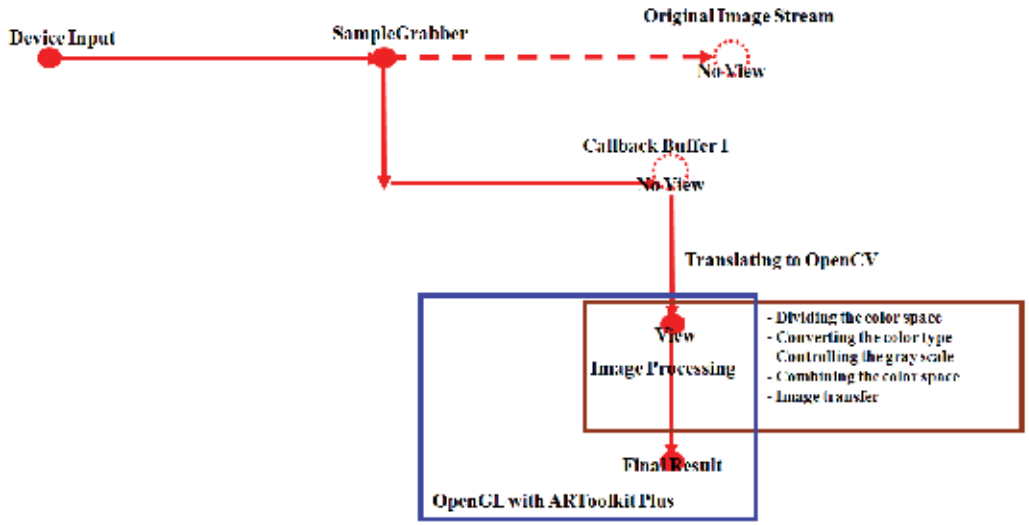


Fig. 7. Image obtaining and conversion steps

However OpenCV does not support digital video image type. To solve this problem, high resolution buffer is obtained by using Microsoft DirectShow technology in this research. The rapid image processing is possible by using various filters. It is a representative advantage of DirectShow. Also well-composed filters are figured out readily through the GraphEdit which is the utility of DirectShow. Based on DirectShow, device filter, preview filter, digital video supporting filter, SampleGrabber filter and Video Renderer filter are used to gather RGB24 image data. Moreover the obtained image buffer is converted into IplImage structure of OpenCV. This technique has the ability to access readily and offers the efficient method which converts RGB24 image data into different image type. The last image type is BGR24.

3.2 Position tracking of target markers

The position tracking of the target marker is the core technology that is determinant for precision and accuracy of AR technology. For tracking target markers, many methods, such as mechanic, magnetic, and optical are examined. Here, the optical method is widespread due to its high precision among the tracking methods. The camera device reads real-time video streams to generate see-through effects on the display equipment. Then edge detection is performed by thresholding with constant value. After this step, the information of detected marker is used to track the local coordinate in the centre of the marker by using matrix calculation.

There are some pose estimation algorithms as the robust planar pose algorithm, the fast pose estimation algorithm, and so on. These algorithms were ported to C++ added to the open source libraries as ARToolkit, ARToolkitPlus, ARTag, and so on (Daniel et al., 2007; Stephen et al., 2008; Bimber et al., 2008; Gerald et al., 2006). In order to be robust system, ARToolkitPlus tracking function that includes improved pose estimation quality with less jitter and improved robustness was used to track coordinate system. The important function for using is rppGetTransMat() (Daniel et al., 2007). This function performs matrix parameters translating to transmit the coordinate data to the rendering module. For this achievement, codes of the matrix parameters translating function and of the tracking function were generated.

The virtual object can be superimposition on the real scene by using each ID-based single marker. Moreover the multi-marker which has one local coordinate system can be used to track. Every single marker size was set 100mm and the user can change the marker size as required at the user interface. The multi-marker size can be changed in the designated data file.

3.3 Implementation of virtual objects loading and superimposition of scenes

The superimposition of scenes is performed by using the obtaining image data, the tracking positions and the virtual objects. The system requirement of the hardware is prime concern to draw the scenes into viewing rectangle of the user interface. Therefore, the overall specification of the hardware as a graphic device, CPU and memory devices have to be high. The high specification of the hardware prevents jitter and lag of the superimposition scenes (Rhee et al., 2007). Also the VRML (Virtual Reality Modeling Language) object files decrease the software overhead and improve the problems of jitter and lag status. All things considered, the object control panel can load text files (.txt) which include the coordinate information of the virtual object based on rearrangement coordinate matrices. And the data

structure including the two-dimensional array was designed for each object rendering on the marker. The data structure has the information of each marker number, visible/invisible, object number and translation-rotation-scale of each object as shown in Fig. 8. These processes are performed by using OpenGL open source graphic library.

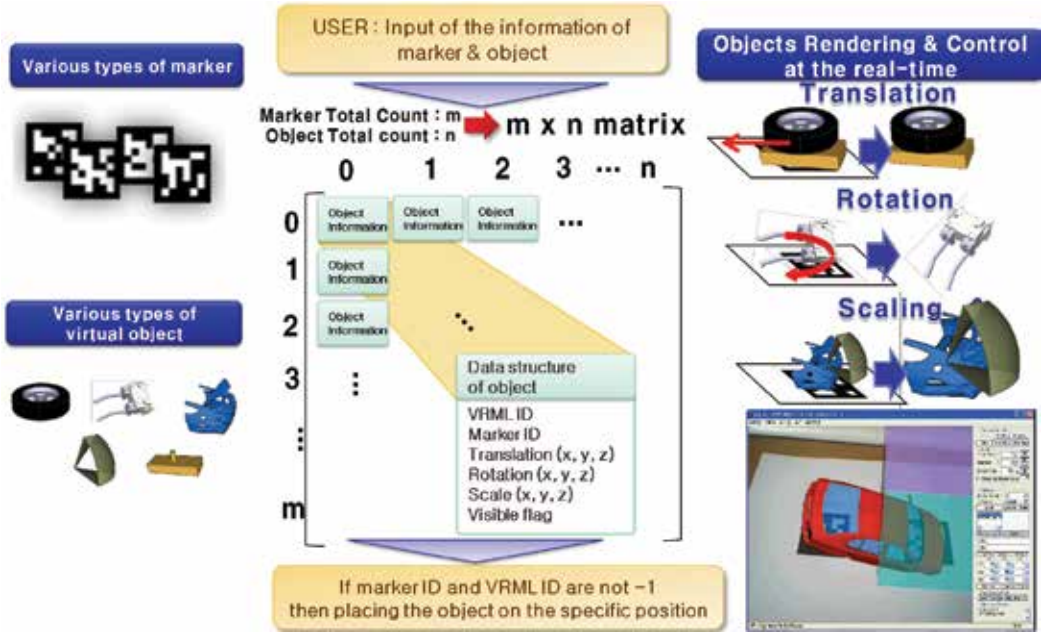


Fig. 8. Matching and control structure of the 3D object information

This information can be controlled through the object control panel. The translation unit value of the virtual object is 1mm and the rotation unit value of the virtual object is 1°. In the case of the scaling, required input value through the user interface is 100 when the side length of the square marker is 100mm. If the side length of the square marker is 40mm, the input value must be 40 to satisfy 1:1 scale between the viewing object and the modeling object. With the information matrix in Rendering module, the matching of the marker and the object and the manipulation of objects are done.

3.4 Implementation of auxiliary tools

The clipping planes can be generated to check the collision between the virtual objects and to ensure the inside area of the virtual objects. The glClipPlane() function of OpenGL was used to generate clipping planes. Therefore, the user can generate 6 planes and can use them at the same time. The normal vector of the clipping plane is derived from the equation of plane to achieve same effect as object controlling. The equation of plane is derived by using one known normal vector and one known point in the three axes generally. However, the normal vector of the clipping plane is derived by using one known point O(0, 0, 0) and three angles which were inputted by the user as shown in Fig. 9 and Eq. (1). And the point O(0, 0, 0) is always used to calculate the equation as the known point.

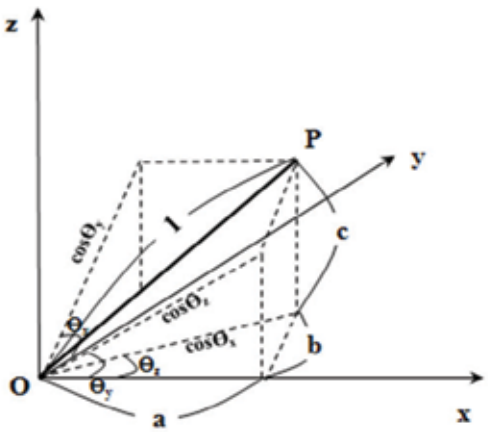


Fig. 9. Relation of known three angles and the normal vector of the plane

$$\begin{aligned} a &= \cos(\theta_y)\cos(\theta_z) \\ b &= \cos(\theta_x)\sin(\theta_z) \\ c &= \cos(\theta_z)\sin(\theta_y) \end{aligned} \tag{1}$$

where, a = vector element of x direction, b = vector element of y direction and c = vector element of z direction.

$\theta_x, \theta_y, \theta_z$ = three known angles

Through the derived normal vector decides the direction of the plane and the plane can be rotated between -180°and 180° by the user's activity. In this case, both the translation and scaling were not considered.

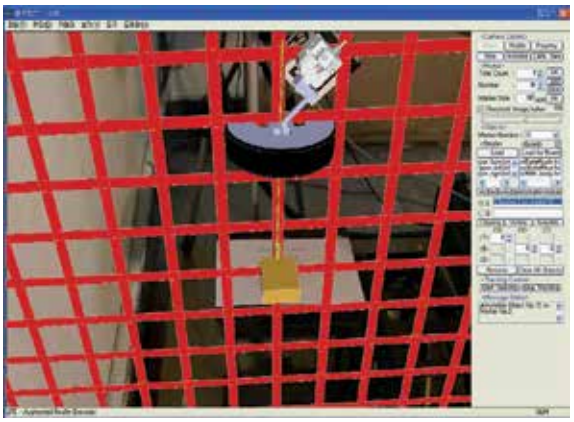


Fig. 10. Using the clipping plane

The generated clipping plane does not have visualization feature. Because of this problem, there is the difficulty to recognize the location of the clipping plane in the virtual space. To solve this problem, the virtual object of the grid plane is created basically and it is matched with the clipping plane at the same time as shown in Fig. 10.

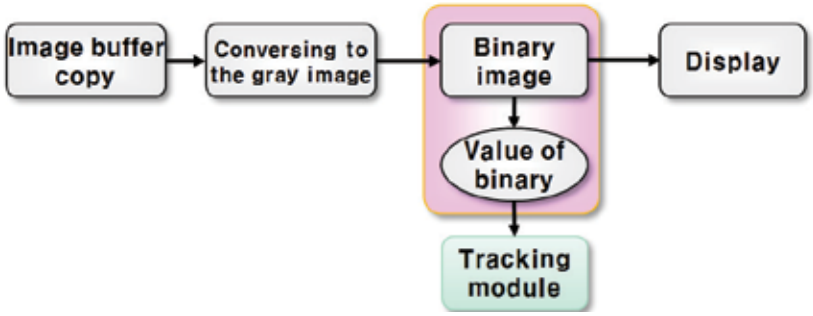


Fig. 11. Binary image value control steps

The manufacturing field has poor light condition. For this reason, the optical marker cannot be recognized easily. To supplement this handicap, threshold value is controlled through the panel. The function of controlling threshold value returns selected pixel value of the binary image from 0 to 255. To control the threshold value, the functions of OpenCV are added as a cvThreshold.13 Binary image value control steps is shown in Fig. 11.

4. Configuration of manufacturing system based on augmented reality

4.1 Component modeling work for system configuration

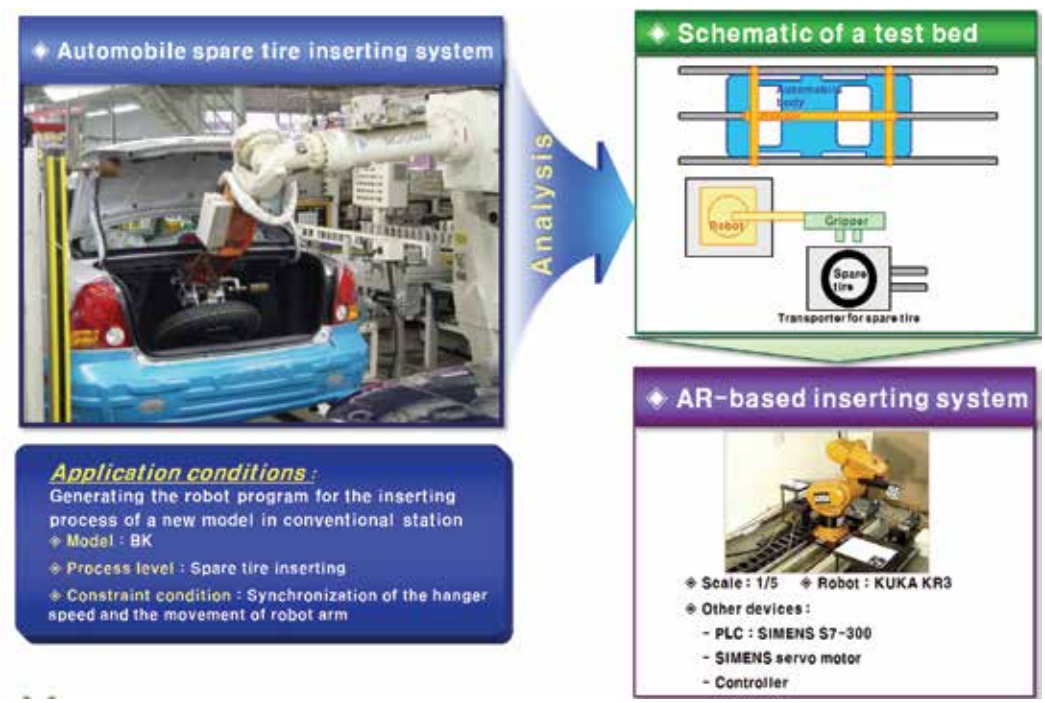


Fig. 12. Development of AR-based assembly system

Through the integration of the previous developed modules, the AR system was implemented. For proving the functionality of the developed AR system, it is applied to a practical problem, i.e. placing spare tire on a mount hole in trunk. The task of AR system is

to generate the robot program of this inserting process for a new model in the conventional station. For this purpose, a test bed with 1/5 scale of a real system was realized(Fig. 12). Three markers and cameras were installed in the appropriate location(Fig. 13). In order to increase work efficiency, it is better to set three display devices which are connected with one PC because each monitor can display the images of each camera. In this case the operator does not need to change each view of the scene.

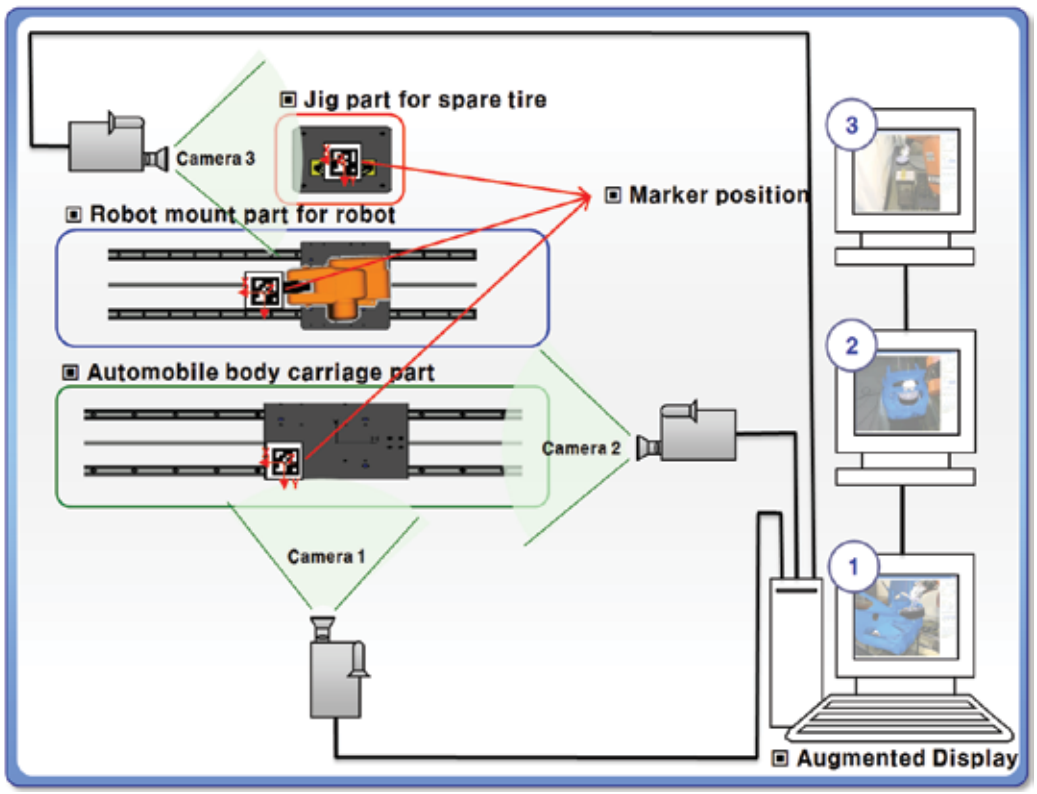


Fig. 13. Installation of cameras for the AR-based inserting system

The developed AR system is applied to the reconfiguration process of the automobile cockpit module assembly system which is the test bed for the real manufacturing system (Park et al., 2008). The target system is the automobile spare tire inserting system. Almost new components have to be remodeled. However, the air finger, the connecting part of the robot and the ball screw are changed partial dimension except overall outer shape. In this case, the existing facility of the test bed is used as it is. The gripper and air finger which perform assembly work, the automobile body for transfer, and the jig/fixture for the automobile body, the encoder and coupling for the synchronous velocity are continuous parts of the virtual object structurally. And these parts perform similar work. Therefore, the each 3D model of them is generated. And the spare tire, the mount for the robot and the jig/fixture of the spare tire are modeled individually. Moreover, the auxiliary tools for collision-free between the peripheral unit and for accurate assembly work were modeled. The centre datum line was generated individually to match between the assembly part of

the automobile body and the spare tire. And the approach path was set and was modeled to avoid collision between the robot and the automobile trunk. The operator is able to perform generating of the robot operation program precisely by using these auxiliary tools. Fig. 14 shows the 3D models for system configuration.

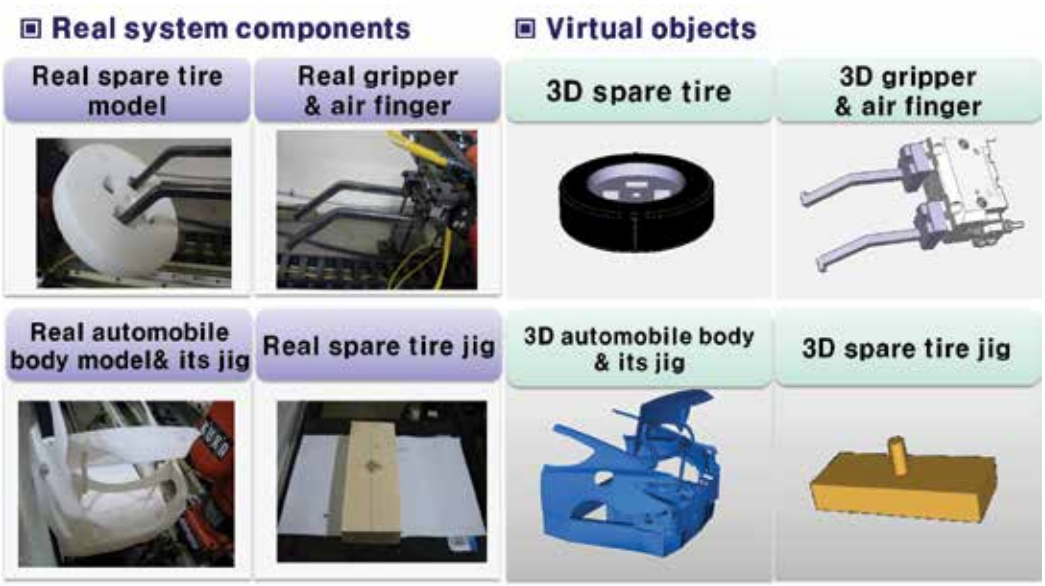


Fig. 14. Preparation of the system components for carrying out the inserting process

4.2 Setting coordinate systems based on markers

Fig. 15 illustrates the generated coordinate systems for configuration of the spare tire inserting system. To perform superimposition of the generated 3D models, the coordinate system for positioning virtual objects has to be established. The Generated 3D models are divided into two groups.

The first group has dynamic motion and the second group does not have movement. The spare tire, the gripper and air finger which perform assembly work require independent one coordinate system (coordinate system No.1). Also the centre datum line of the spare tire for assembling serviceability used the coordinate system No.1. The coordinate system No.1 has to be mounted on the end of the robot gripper. Moreover, the system should be seen in every direction due to free movement of the robot. Therefore, the coordinate system No. 1 which has several markers was manufactured in the structure of a box type. And the automobile body and its jig/fixture on the carriage part require independent one coordinate system (coordinate system No. 2), because the carriage plate influences to the automobile body and its jig/fixture. The centre datum line of the automobile body and the path of assembly work used the coordinate system No. 2. In addition, one marker which considers the position of the camera device was established. The encoder, the coupling, the mount for the robot and the jig/fixture of the spare tire are fixed components and they have collision-free during assembly work. So these components use one coordinate system (coordinate system No. 3).

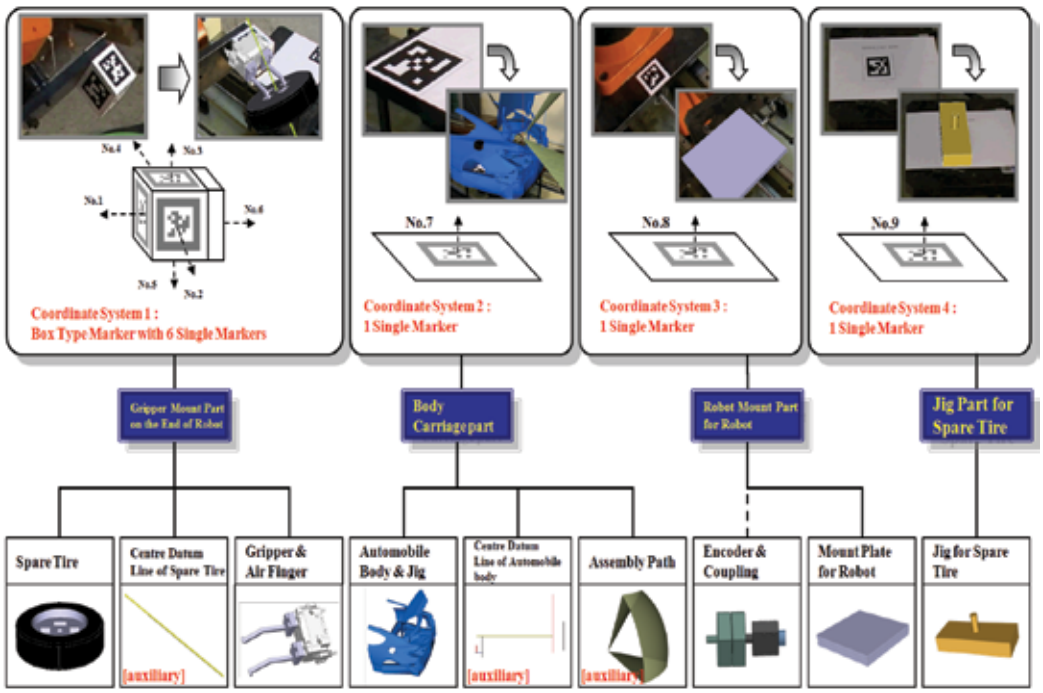


Fig. 15. Design of the markers for establishing the coordinate systems

4.3 Generating operation program for spare tire inserting system

The test bed of the automobile spare tire inserting system is configured as a 5X-reduced scale. The operator generates the robot operation program with two cameras. One camera is set at the side of the test bed and the other is set at the behind of the test bed.

The clipping plane is generated for the alignment of the inserting objects, i.e. spare tire and mounting hole and for calculating the moving distance of the spare tire. For the collision free inserting of spare tire, the auxiliary model guides the inserting path. With the previous mentioned installation, a programmer generate the robot program for the inserting process with teach pendent in front of a monitor, i.e. the programmer carry out the programming process without seeing the real system. Through down loading the generated robot program, the correctness of it can be proven. The completeness of the generated operation program was proven by applying it to the conventional assembly station. The boundary conditions, such as the position tolerance range within 1~5 mm, were fulfilled. Fig. 16 illustrates the generating operation program of the automobile spare tire inserting system by using AR.

5. Conclusion

The AR system which can be used at the manufacturing system was developed based on general software development method. The fundamental functions as the video interface, the tracking and rendering were implemented. And the manufacturing field condition-oriented auxiliary functions were added.

Fig. 16. Generating a robot program for operating the insert processes

The function of generating clipping plane and controlling for precision and accurate work was used effectively during generating robot operation program. And the threshold value control function has similar effectiveness. These results attest functionality of the developed AR system which has high usability when the manufacturing system configuration is performed.

The benefits of the developed AR system read as follow:

- Generating an operation program of a manufacturing system for new model without prototypes and in a conventional manufacturing system.
- Modification of the conventional manufacturing system for manufacturing new models in advance.
- Reduction of ramp-up time dramatically.

The module for measuring the distance between virtual objects in virtual world will be developed. Moreover the module for simulating dynamic behaviors of virtual objects will be developed. And for applying to the mass production system, CCD camera should be applied to acquire images. In the developed AR system, digital video camcorder is used.

6. References

Daniel, W. and Dieter, S. (2007). ARToolKitPlus for Pose Tracking on Mobile Devices, *Computer Vision Winter Workshop*, pp. 1-8.

Stephen, C. and Mark, F. (2008). *Augmented Reality a Practical Guide*, Pragmatic Bookshelf, pp. 117-153., ISBN 1934356034, USA

Bimber, O. and Raskar, R. (2008). *Spatial Augmented Reality - Merging Real and Virtual Worlds*, A K Peters, pp.1-12., ISBN 1568812302, USA.

Ong, S. K. and Nee, A. Y. C. (2003). *Virtual and Augmented Reality Applications in* Manufacturing, Springer, ISBN 1852337966, USA.

Günter, W. and Emmerich, S. (2005). Digital Planning Validation in Automotive Industry, *Computers in Industry*, Vol. 56, pp. 393-405.

Park, H. S. and Choi, H. W. (2008). Development of a Modular Structure-based Changeable Manufacturing System with High Adaptability, *International Journal of Precision Engineering and Manufacturing*, Vol. 9, No. 3, pp. 7-12.

Kim, S. C. and Choi, K. H. (2000). Development of Flexible Manufacturing System Using Virtual Manufacturing Paradigm, *International Journal of Precision Engineering and Manufacturing*, Vol. 1, No. 1, pp. 84-90.

Park, H. S. and Lee, G.B. (2007). Development of Digital Laser Welding System for Automobile Side Panels, *International Journal of Automotive Technology*, Vol. 8, No. 1, pp. 83-91.

Park, H. S. and Lee, H. B. (2006). Devlopment of Digital Assembly Cell for Laser Welding of Side Panels, *International Journal of Automotive Technology*, Vol. 7, No. 6, pp. 721-728.

Lee, K. H., Lee, J. M., Kim, D. G, Han, Y. S. and Lee, J. J. (2008). Development Technology of Vision Based Augmented Reality for the Maintenance of Products, *Transactions of the Society of CAD/CAM Engineers in Korea*, Vol.13, No.4, pp.265-272.

Park, H. S., Choi, H. W. and Park, J. W. (2008). Implementation of AR based Assembly System for Car C/Pad Assembly, *Journal of the Korean Society for Precision Engineering*, Vol.25, No.8, pp.37-44.

Gerald, S. and Axel, P. (2006). Robust Pose Estimation from a Planar Target, *IEEE Transactions on pattern analysis and machine intelligence*, Vol.28, No. 12, pp. 2024-2030.

Rhee, G. W., Seo, D. W. and Lee, J. Y. (2007). Ubiquitous Car Maintenance Services Using Augmented Reality and Context Awareness, *Transactions of the Society of CAD/CAM Engineers in Korea*, Vol.12, No.3, pp.171-181.

16

Autonomous Evolutionary Algorithm

Matej Šprogar
Faculty of Electrical Engineering and Computer Science
University of Maribor
Slovenia

1. Introduction

Evolutionary algorithms (*EA*) are randomized heuristic search methods based on the principles of natural evolution (Banzhaf et al., 1998; Goldberg, 1989; Holland, 1975; Bäck, 1996; Koza, 1992). If we know how to describe the problem using the terminology of artificial evolution, the EAs are quite easy to apply. Actually, the search for solution(s) is *transformed* into a search for the best EA setup – a mixture of highly correlated settings and functions (encoding scheme, run-time parameters, fitness (objective) function, selection mechanism. . .). Finding a good EA setup is a problem because EAs are chaotic systems where small variations in initial setup produce large variations in the long-term behavior of the model. A good setup for one problem is mostly unusable for another, although similar problem.
Evolutionary algorithm that would be easy to apply in any problem domain would have to be autonomous in a sense that it would regulate its own behavior and would have no need for human intervention (except for the preparation phase, of course). This article discusses the operating principles of such an algorithm and presents its implementation. The *Autonomous EA* (AEA) is an experiment in the evolution of evolutionary algorithms. It is not much different from existing EAs and the line between the two is sometimes very blurred. Actually, AEA combines known concepts, insights and solutions from EAs, artificial life, chaos theory and complex adaptive systems theory into a new form of evolutionary algorithm.
The nomenclature used in different fields is overlapping (for example individual/ solution/object/agent). In AEA the evolving individual represents the solution: a population of individuals (solutions) is evolved in order to find a solution (individual) for the problem at hand. Population is just a limited representation of the vast search space of all possible solutions.

1.1 Controlling evolution

Evolutionary computing is an artificial world where computer-based models are directly written in terms of conditional actions and operations. These models can then be "run" in a simulator. Like any other, the EA simulation is controlled by many parameters. There are numerous studies of EA parameters giving suggestions on their "correct" values. Different types of control of algorithm parameters can be classified as either parameter tuning, deterministic parameter control, adaptive or self-adaptive parameter control (Eiben et al., 1999).

The EA implementers always try to create the system with as little parameters as possible. To reduce this already minimal number of parameters even further, two possibilities exist:

1. make the parameter fully self-adaptive; or
2. remove the need for a parameter.

Self-adaptation effectively "hides" a parameter and can be used for most, but not all parameters. Because self-adaptation is a much explored topic (Angeline, 1995; Eiben et al., 1999), we chose to investigate the second possibility more. We can try to remove a parameter either by replacing it with some non-parametric value/principle or by changing the operating principle(s) of the EA. Certain parameters, however, cannot be removed – for example the hardware limitations (amount of memory, available CPU time. . .).

EAs are complex systems with numerous algorithm parameters to set. For example, Genetic Programming (GP) in the Open Beagle evolutionary framework by default already includes over 20 different parameters that directly control the behavior of the underlying evolutionary computation (*ec.pop.size*, *ec.repro.prob*, *ec.sel.tournsize*, . . .) (Gagne & Parizeau, 2006).

1.2 Objective

The objective of Autonomous EA is to decrease the number of algorithm parameters to a bare minimum. The population size, for example, is one problematic parameter despite numerous attempts and claims of the 'optimal' value. Next problematic parameter is the one that controls the termination criterion. The most troublesome, however, is the fitness function. Although the fitness function does not seem to be a parameter, it is a rather complex mixture of criteria of unknown/changing importance. It must be "input" to all existing EAs by the human operator, therefore it is a parameter. Fitness can not be self-adapted (only the constraint weights of fitness function can be self-adapted) or self-induced as EAs will quickly learn to "cheat" by producing worthless individuals with good fitness. Fitness is fundamental and is actually a reason for many other control settings.

2. Fitness – the core of the problem

Nature is using simple atomic rules to guide the evolution. Paradoxically, canonical EAs are already too complicated to follow this simple rule. They constantly apply the same orthodox idea of fitness and associated complex mechanisms in their evolutionary loops. Small mistake in fitness and EA will fail to find an otherwise obvious solution altogether. Creation of adequate fitness functions demands significant knowledge of the environment to be evaluated – this can imply that the problem might have already been solved or that other, non evolutionary technique(s) might be more efficient (Angeline & Pollack, 1994). Inadequate fitness makes EA focus on sub-optimal solutions. To avoid local optima different "corrections" and "tweaks" are usually applied. These corrections need further corrections resulting in complex EAs. In EAs, fitness is at the core of everything.

Fitness function is the engineer's interpretation of the problem and is as such affected by the computational biases of human cognition (Stanovich, 2003). Biased EA is unable to discover the *generalists* – solutions with the capability to generalize – because generalization can only be the (emergent) property of solutions produced by the unbiased EA. Solutions produced by biased EA are on the contrary often brittle – they fail on the previously unseen/new data. Unless EA is used to find *specialists* (optimizations of a single fitness peak), this is a major problem. Solution is either to make the fitness unbiased or to "remove" it altogether.

2.1 The search space

The success of search depends mostly on the distribution of solutions within the search space and this distribution is determined/described by fitness. For the same simple problem a perfect fitness, for example, defines a gradual landscape well suited for deterministic methods like gradient descent; sub-optimal fitness defines a more complicated space with many spikes; and bad fitness creates a chaotic space where any informed search is impossible – in such space a random search is as successful as any other search (Wolpert & Macready, 1997).

Existing fitness-based EAs are efficient in beautiful fitness problem spaces. Differential Evolution (DE), for example, is an effective, simple and fast optimization algorithm (Price & Storn, 1997). However, EAs are mostly tested on clean mathematical problems where fitness *optimization* is the problem. More difficult learning problems (for example data mining or genetic programming) require searching in unknown spaces for solutions of unknown structures – there, fitness *itself* is the problem (at least the fitness in a form needed by EAs).

2.2 The essence of fitness

The idea behind fitness is to produce a single number that will (magically) create a hierarchy of candidate solutions – the individual at the top *better* than the ones underneath. The tendency to over-simplify by squeezing into one single fitness number matters that are too rich to be described by it is a typical human mistake (an example is stock market analysis (Mandelbrot & Hudson, 2004)). It results in a fragile system. In EAs the concept of fitness *looks* simple and because all other mechanisms are crafted to suit the concept of fitness, the resulting system is extremely complex and not at all simple. In fact, the complexity of EAs is reflected in the number of necessary algorithm parameters.

In the real world, there is no *ultimate hierarchization* of individuals nor are individuals living under the same conditions. Darwin's survival of the fittest should not be interpreted as "evaluate, pick the best and kill the rest", it should be a "live and let die and the best will prevail with more offspring" approach. This way the butterfly-effect the discarded (*presumably* insignificant) solutions might have had on the evolutionary process would not be lost!

The dictionary defines Darwinian fitness as the number of offspring or close kin that survive to reproductive age (Dictionary.com, 2006). This definition is impossible to directly implement in EAs because artificial individuals do not live a life. Fitness is a *shortcut* that allows EAs to by-pass the phase of life of individuals. Fitness effectively determines the number of surviving offspring and replaces the individual's true Darwinian fitness! This shortcut only seemingly produces a desired result because life is a chaotic process, extremely susceptible to initial conditions. Ignorance of this results in many problems associated with EAs (Toffoli, 2000).

2.3 Limitations and assumptions

The EAs are always limited by the availability of computational resources. Bremermann states that faster computers are insufficient, "we must look for quality, for refinements, for tricks, for every ingenuity that we can think of." (Bremermann, 1962). This is why the EA community regards early Friedberg's approach (Friedberg, 1958; Friedberg et al., 1959) as being immature, attributing this mostly to the fact, that he used the so-called *binary* fitness. EAs are going in the direction of establishing more elaborate fitness measures and selection schemes. The argument for this is the belief that without an accurate enough ranking, the

natural dynamics of artificial evolution might be compromised (Banzhaf et al., 1998; Angeline & Pollack, 1994). If unlimited computational resources were available, EAs could operate without discarding any individuals. Because this is not the case, EAs must employ some sort of artificial selection, selection being "a name for the ability of those individuals that have outlasted the struggle for existence to bring their genetic information to the next generation" (Bäck, 1996). The EAs fail to follow this simple definition because they falsely assume that:

1. evolution is a process of fitness-based ordering and selection; and
2. the individual's fitness is measurable and is independent of the fitness of its offspring.

Firstly, individual's fitness is based on the observed performance and EAs use this score to justify the selection, although biologists argue that "fitness cannot be used as a cause but merely as a description of natural selection" (Henle, 1991). EAs mistake the measurement of ability to fit the purpose for survival. Interestingly, EA researchers are aware of the fact that biological struggle for existence has no counterpart in EAs, yet they ignore this or find it at most an interesting research field (Bäck, 1996).

Secondly, EAs calculate the fitness (f) without considering the individual's potential to have better offspring. EAs rely on false impression that only fitness for purpose is important. In fact, the *ability* to create good offspring and forward the genetic information into next generations is what survival is about, yet this can not be computed in advance. If one were trying to compute it anyway, fitness values of all of individual's offspring would be needed and for that the fitness values of all of the offspring's offspring would be needed, etc.:

$$f(x) = u(x) + f(\hat{x}), \tag{1}$$

where u is a performance/utility function (for example accuracy) measuring the success of x at solving the given problem and $\hat{x}$ is x's offspring ($f(\hat{x}) = \sum f(\hat{x}_i)$). This is a highly recursive definition. Canonical EAs assume that the individual with better fitness will produce better offspring than the competing individual with worse fitness score, or at least that the probability for the opposite is low enough:

$$P(f(\hat{y}) > f(\hat{x}) \mid u(x) > u(y)) \approx 0\,, \tag{2}$$

A basic assumption of the fitness function, namely that seemingly-better individual is assigned a better (in example higher) score, neglects the fact that the seemingly-worse individual can posses the much-needed building block of the global solution. The search space is normally so enormous that EA must not afford to lose this building block although located in a seemingly bad individual. But because of the limited computational resources, EAs use

$$u(x) > u(y) \Rightarrow f(x) > f(y)\,, \tag{3}$$

to simplify (1) into

$$f(x) \approx u(x)\,. \tag{4}$$

For fitness calculation standard EAs rely on (4) because it's easy to implement and execute. The problem is that (4) does not link the ability to survive with the ability to solve a problem in any way, whereas (1) automatically makes this link. In standard EAs, survival is

artificially directed by the *selection* of survivors. Eq. (4) makes the evolution discover local solutions because of premature convergence problems. EAs acknowledge this and try to compensate by applying special mechanisms, which again have side-effects for which additional fixes are needed. . . In EA literature, numerous statements like: "the niching methods have been developed to reduce the effect of genetic drift resulting from the selection operator in the standard GA (Sareni & Krähenbühl, 1998)" can be found.
Something similar occurs, for example, when evaluating the chess board position. The quality of a player's move is determined by evaluation of the board after the move. If perfect fitness were available, the computer would *always* beat the human champion. Even worse, it could tell the winner right from the start! However, perfect fitness would only be possible to compute if fitness values of *all* successive moves were available; for these the scores of all the successive's successive moves would be needed. . . In computer chess the effort has not gone to teaching computers about chess, but to improving the algorithms for deciding when to cut off calculations and when to calculate more deeply. Something similar occurs in EAs, where special fitness-based mechanisms (e.g. different criteria weights, selection schemes . . .) are introduced and "improved" all the time, but fitness as a concept is never questioned.

2.4 Implicit perspective

The quality and quantity of population's members are coupled properties – quality affects quantity and vice versa. Holland described this phenomena in the terms of adaptive complex systems (*cas*) (Holland, 1995). He, like many others, recognized that fitness must "depend on the context provided by the site". Unlike EAs, however, he placed particular emphasis on avoiding an overt fitness criterion. He introduced the concept of a *resource* and his agent could reproduce only after it had acquired enough resources to make a copy of itself. Holland effectively replaced traditional explicit fitness with fully implicit resource acquisition. He avoided the fitness calculation because fitness is implicitly defined by the resource acquisition of his agents, which live or die in terms of their ability to collect critical resources.
Holland's principles were never successfully applied to engineering problems (where emphasis is put on finding *best* solutions). Rather, they've been used for studying complex adaptive systems, natural systems and in Artificial Life (AL) research. Artificial evolution in engineering was always based on explicit fitness; the numerous constraints implied by engineering problems make application of Holland's ideas troublesome because they're not directly goal driven as the engineer needs it.

2.4.1 Co-evolution

Co-evolutionary search should be more successful than 'complete' static fitness evaluation because co-evolving individuals sample the problem space more efficiently (Angeline & Pollack, 1994; Pagie & Mitchell, 2002; Hillis, 1990). Paredis, however, observed that in some cases co-evolution does not lead to better results (Paredis, 1997). These cases are often characterized by the occurrence of the so-called *Red Queen dynamics*[1] (Pagie & Hogeweg, 2000; Juille & Pollack, 2000), which can be prevented from persisting by the heterogeneity in the populations (Pagie & Mitchell, 2002).

[1] Evolutionary change may be required to stay in the same place. Cessation of change may result in extinction.

Co-evolutionary search in EAs is mostly said to use "implicit" fitness, where the fitness of evolving solutions depends on the state of other, co-evolving individuals. This, however, is not implicit according to Holland's definition because co-evolution is still calculating and using fitness. Fitness must not be a calculated value but rather an observed property of an individual; this observation can only be made *after* the individual has performed actions in its world (lived a life)!

3. Autonomous Evolutionary Algorithm

To achieve autonomy from human intervention AEA employs co-evolution of two competing populations. The fight for survival between two individuals, one from each respective population, simulates life and determines survivors in the simplest and most unbiased manner possible; in the process the number of offspring and individuals to be discarded are determined (this is necessary to keep the simulation within the available memory limits). The co-evolution terminates automatically[2] after one population dominates the other.

Standard EAs use fitness to create a hierarchy (ranking) of individuals. Position within this hierarchy defines the number of offspring. AEA, on the contrary, does not rank the individuals. Rather, it mimics a predator-prey like system, where individual survives only by outperforming another individual. In the process an individual holds and collects a virtual resource – energy – which is needed to create and shape the offspring. AEA essentially simulates the *flow* of energy between the two co-evolving populations.

3.1 Life of an individual

Standard EAs treat individuals as non-living objects. Individuals are created, evaluated and very likely also immediately destroyed without any impact whatsoever on the rest of the population. AEA, on the contrary, makes each individual alive – each living individual has to fight for survival and only surviving individuals reproduce.

AEA maintains two separate populations of individuals of the same type. Fight for survival is a simple competition between two randomly chosen individuals, one from each respective population. The competition is about solving an atomic task from the problem at hand. The better of the two competing individuals at this atomic task is the survivor. By using an atomic task we make this process as unbiased as possible[3].

AEA never tries to judge *how-much-better* one individual is compared to another. By making every individual "alive" (putting him through the fight-for-survival test) we get a list of survivors, what effectively removes the need for a selection phase.

The loser's energy reserves are transferred to the winner and then the loser is removed from the system – the allocated spatial resources (every individual occupies a certain amount of available virtual space / physical memory) are again made available for offspring. The energy determines the creation (number and genetic structure) of offspring. The main result of individual's life are fluctuations in quantity and distribution of virtual energy within the AEA system.

---

[2] Of course the evolution run can also be terminated artificially.

[3] Typical atomic task is one fitness case from the learning database.

3.1.1 Compared to traditional EA

The regular EAs look similar – the selection operator can give an individual zero, one or even more chances to "survive" and reproduce (this number is either random (e.g. tournament selection) or based on the fitness (e.g. roulette-wheel selection)). AEA, on the contrary, selects individuals to be removed only *after* they participated in the fight-for survival and lost their energy to the winning individuals. Effectively this is a "select all" selection strategy followed by a necessary[4] removal of some individuals.

Next important difference is how the fight for survival is made. To make the outcome as unbiased as possible, AEA employs the stochastic sampling of a *single* problem instance (atomic task) as the minimal measure of competence. Consequently also an overall bad performer can win against the almost perfect opponent, especially if the learning data contains noise. AEA lets the evolution decide which individual is better in the long run. . .

AEA goes the opposite way of traditional fitness. Instead of using all available information, AEA makes use of only the tiniest fraction of it. AEA does not give importance to how many problem instances does one solution solve nor does it make any biased presumptions whether it is better to solve one instance over another. It just says: for this atomic task, this solution is better. The worry of handling noisy or missing data is left to the evolution. Sometimes a good individual will be defeated by a weak one; good individuals, however, have more energy and more offspring and will therefore probably survive at other opportunities.

3.2 The core of the autonomous algorithm

The main resource needed for reproduction is energy, which is exchanged only between competing individuals. The winning individual simply collects the loser's energy reserves. Second resource in AEA is the space – each individual occupies a certain amount of memory. Each gene in individual's genotype occupies one unit of memory space – size of an individual equals to number of its genes. Both populations have limited space for holding individuals. After the population space is full, offspring production is suspended until further individuals are removed from the population.

The algorithm 1 shows AEA's core. At start, the two co-evolving populations P_1 and P_2 are created and randomly initialized to fully occupy the assigned memory space. Individuals $x \in P_1$ and $y \in P_2$ are fully qualified solutions. The initial sizes of populations ($|P_1|$, $|P_2|$) depend on the average size of fresh individuals created by a typical initialization routine. Individuals should live simultaneously but because today's computers employ serial CPUs we can only simulate this parallelism. First, random interaction pairs (x,y) are determined using the *random_pairs* function, which selects two random, previously unprocessed members, one from P_1 and the second from P_2, respectively. This creates a set of random pairs Q. Because P_1 and P_2 in general differ in size ($|P_1| \neq |P_2|$) only the smaller of the two populations is fully used in one cycle of the main loop ($|Q| = min(|P_1|,|P_2|)$); the remaining individuals will be processed in the next cycle(s)[5].

Next is the actual life of an individual: from each available pair $(x,y) \in Q$ the winner is determined. Winner takes the loser's energy and waits for breeding, while the loser is discarded in order to free its memory space in its native population. Fight for survival redistributes the energy and frees the spatial resources.

---

[4] Limited computational resources require a limited population.

[5] This must be guaranteed by the *random_pairs* function.

Ideally, the surviving (active) individual would get one chance to breed and produce active offspring. This is unfortunately impossible to implement because the offspring would quickly overfill the population's memory space. To avoid this AEA employs a simple waiting list (*children*), where the inactive offspring waits to be included in the active population. Until the list is emptied no further breeding takes place. Of course, other workarounds are possible. Before the main cycle is repeated and only after all of population's members participated in a fight for survival, the *breed* function produces offspring for respective population (Algorithm 2). New children are produced only if all previous children were processed by the main loop. Function *create_child* produces one child using traditional variation operators (sexual crossover and mutation). The number of inactive children that are transferred into active population depends on the number of free spatial resources. This is to ensure as parallel evolution of all individuals as possible. The size of the resulting population is *not* calculated nor artificially maintained at a certain level as is common in EAs; it is only limited by the available memory space.

Algorithm 1 The core loop of AEA.

```
// two competing populations
population P1(memory/2), P2(memory/2);

while ( true ) {
    // create a set of random pairs of individuals
    // not all individuals are necessarily paired
    Q = random_pairs( P1, P2 );

    // fight-for-survival
    for-each ( pair ( x, y ) in Q ) {
        if ( x wins against y ) {
            x.energy += y.energy;
            P2.erase( y );
        } else {
            y.energy += x.energy;
            P1.erase( x );
        }
    }

    // termination criteria: empty population
    if ( P1.empty() ) return P2;
    if ( P2.empty() ) return P1;

    // only breed the fully processed population
    if ( P1 has been fully processed ) P1.breed();
    if ( P2 has been fully processed ) P2.breed();
}
```

Algorithm 2 Breeding of individuals within population.

```
population::breed()
{
    if ( children.empty() ) {
        // create offspring
        for-each ( individual x in this population ) {
            e = x.energy;
            while ( e>0 ) {
                individual y = random_member( this population );

                child = crossover( x, y );
                child.mutate( e );
                child.energy = 1;

                children.enque( child );
                e = e-1;
            }
            x.energy = x.energy / 2;
        }
    }
    // try to move children into active population
    this population.transfer( children );
}
```

The new-born individuals have by default exactly one energy point; this allows the system to grow and sustain larger individuals because energy is a vital resource in reproduction. The number of descendants equals the leading parent's energy. The higher the collected energy the more offspring the individual has. Each individual is a result of sexual reproduction of the primary parent with a random partner. The main parent produces a child by investing a certain portion (reproductive energy – e) of its energy into creation of the child's genotype. The amount of energy invested directs the creation phase, in particular the selection of mutation point and the probability of mutation of that point. The completion of the breeding phase also takes away 1/2 of the main parent's energy. The energy of the partner does not influence creation of main parent's child; partner only provides its genotype for copying.

3.2.1 Energy and reproduction

The individual's energy level determines:

1. the number of offspring,
2. the mutation probability,
3. the mutation point.

Crossover and mutations are interdependent: a new child is constructed from copies of both parents' genes. Size of the child directly depends on the selection of the random crossover point within each parent. For a duplication of 1 gene 1 reproductive energy point must be provided. When this energy is exhausted, the gene about to be copied is subject to mutation with probability $mut\_prob = \phi/(1.0 + e)$, where ϕ is a random number from interval [0,1]. For

linear genomes the mutation point is therefore the e^{th} gene in the sequence. Tree-like genomes can be copied in either breadth-first or depth-first order. The remaining genes are copied without further energy consumption. The same result can also be achieved by making a mutation-free child and then mutating its e^{th} gene – the procedure used in Algorithms 2 and 3.

Algorithm 3 Mutation depends on available energy.

```
void individual::mutate( e )
{
    mut_prob = φ/(1+e);
    if ( random(mut_prob) ) {
        if (e < |genome|) {
            gene = genome[e];
            mutate( gene );
        }
    }
}
```

The parent's energy is not always used in full when making children. Rather, children are created with decreasing amounts of energy. This way a constantly changing number and position of possible mutations is introduced in offspring – the first child's reproductive energy *e* equals parent's energy, the second child's reproductive energy is one less etc. A "good" parent has enough energy to copy most of its genetic material faithfully and produce offspring with little or no mutations. The children of larger individuals without large energy reserves, however, are more likely to undergo mutations.
If *energy* > |*genome*| no mutation is performed (as it is impossible to mutate a non-existing gene!). This is advantageous in creating offspring of parents with high energy status with respect to their size – higher energy suggests the parent is successful thus its genes should be preserved. Reproduction code in AEA (Alg. 2) makes pressure for many "good" copies of energy-full individuals. Parents with low energy reserves will have few and less-similar children.

3.3 Example of reproduction

For illustration of the reproduction consider Fig. 1, where a crossover of two GP-trees is shown: the first 5-node parent ("1 + (7 – 4)") holds *energy* = 8 energy points; for the creation

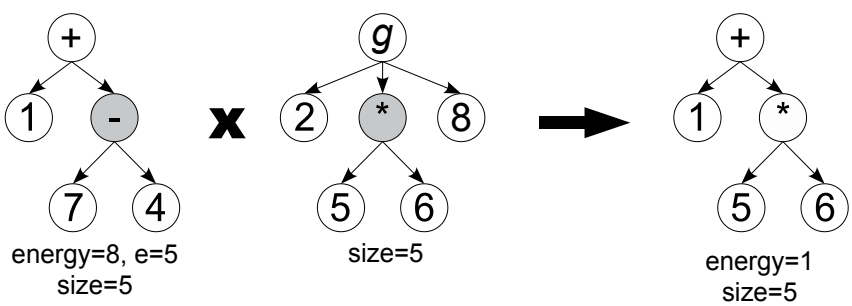


Fig. 1. Crossover without mutations.

of the fourth child $e = 5$ energy points are designated. The partner individual is "$g(2, 5 * 6, 8)$". The randomly chosen crossover points in both parents are '-' and '*'. Regular crossover results in a 5 node expression "$1 + (5 * 6)$". Because $e \geq 5$ (size of the child), all nodes are copied/created without mutations.

If the crossover point in Fig. 2 was selected at '4' then regular EA crossover would produce the expression "$1 + (7 - (5 * 6))$" with 7 nodes. Because there were only ($e = 5$) energy points available for creating this child, AEA is unable to reproduce all 7 nodes. Instead the first five nodes are copied ('+', '1', '-', '7' and '*') and the sixth node is mutated with probability $\phi/(1 + 5)$ from '5' $\rightarrow$ '3'; the remaining node(s) are copied without mutations.

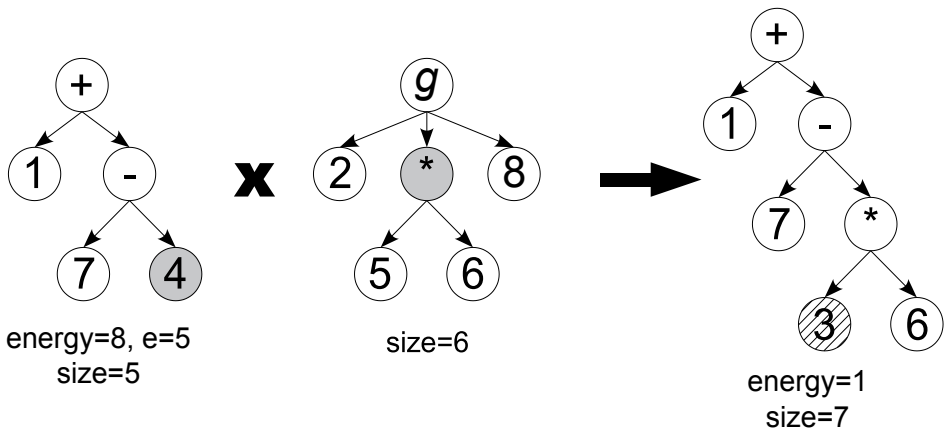


Fig. 2. Crossover with mutations.

3.4 Population size

Autonomous EA does not have any population size setting. This parameter is made obsolete by the idea of co-evolving populations in a spatially limited environment, where energy is used to direct the reproduction phase. An upper limit to population size must exist (or else population's size would explode in a matter of few generations); AEA implements this limitation through a total number of genes a population can hold. Consequently, a population can hold a large number of small individuals or a small number of very large individuals. One gene is said to occupy one unit of computer memory. Available memory is the first unavoidable setting of AEA.

The phase plot in Fig. 3 displays three possible scenarios of how the populations' sizes change during evolution. The evolution generally starts in the point S_0 and progresses through S_1 to terminate in either S_2 or S_3. The area of S_1 is an ever-changing state, where small improvements or changes in P_1 are counterbalanced by changes in P_2 and vice versa. From state S_1 the AEA can escape into either S_2 or S_3. However, if AEA is unable to break out of S_1 in "reasonable" time, it must be terminated artificially.

If one population is initialized to contain "much" better individuals the system will be unable to reach S_1. Instead, the weaker population will vanish too quickly and the system will follow dotted arrows in Fig. 3. The evolution *needs* to cycle in S_1 for some time before any significant progress can be expected. The prerequisite for this is the balanced quality of initial two populations. Only two populations of the roughly same quality level can obey the interaction principles from Fig. 4.

Figure 4 depicts the influences the two populations' quality and size have on each other. Solid arrows represent positive influence and dashed arrows represent the negative impact of one entity on another. The "quality" is impossible to define (or else we'd have a perfect fitness function!) but is a property that should be maximized. This goal is achieved by the positive reinforcing loop: higher quality population will probably remove "low-quality" members from the lower quality population resulting in improvement of the average "quality" and reduction in size of the weaker population. The influence of increase in size on the opposing population's size and quality is sometimes positive and sometimes negative – many trivial solutions can be beneficial for the opposing population, large number of good solutions, however, can be catastrophic.

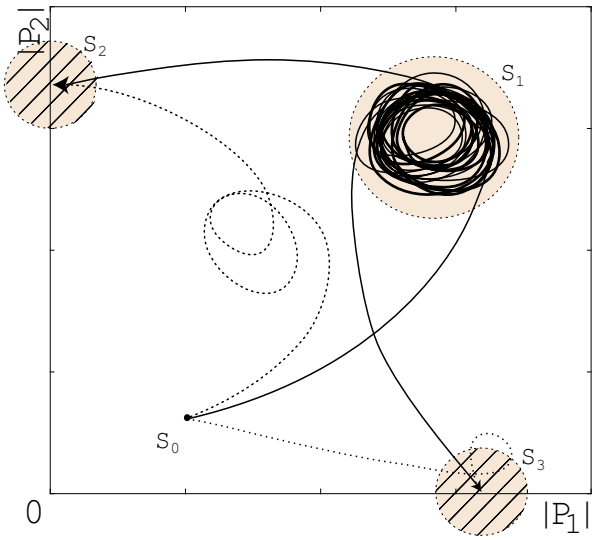


Fig. 3. Typical evolutionary scenarios.

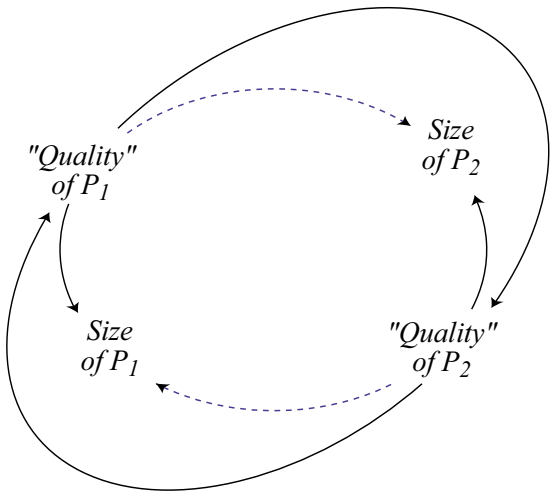


Fig. 4. Relations between quality and number of individuals.

This picture displays the complex interactions that drive the evolutionary search. The link between size and quality is dynamic. In the much studied predator-prey model (Blanchard et al., 1997), only the size of both populations is relevant while the quality by definition remains constant. In AEA the quality is expected to vary making this model more complex, though it should exhibit behavior similar to that of the predator-prey model.

3.5 Run termination in AEA

In standard EAs, run termination seems a non-problematic parameter: the EA run simply terminates if fitness hits a pre-set threshold or after a certain amount of processing has been done. The problem is again fitness. As discussed before, fitness is not an objective measure of success and can thus not be used as a termination criterion (see also section 4). The processing time, unfortunately, must remain a mandatory parameter for AEA, too.

Natural populations live in an ever-changing environment (S_1 area), where they're constantly challenged to improve their qualities. In history, the species became extinct for various reasons (for example the meteorite (supposedly) wiped out the dinosaurs, humans killed the dodo birds. . .). AEA treats one population as a species fighting for supremacy over another species. Because the two populations (P_1 and P_2) are genetically fully isolated, they can physically represent the same but "logically" different species. The main auto-termination criterion for the main loop is therefore the moment of absolute victory – the $|P_i| \rightarrow 0$ moment.

If there's no such event for a predefined amount of time, the evolution run can be interrupted artificially (just like in standard EAs). The problem remains, however, how to recognize/select the "best" solution from the remaining individuals. In traditional EAs, fitness-best individuals are proclaimed general solutions, yet there is no fitness criterion integrated within the AEA. One option is to use the classical fitness function just to select the resulting individual from the final population. This fitness is *not* used to guide the evolution in any way; it is rather calculated only after the evolution has already (been) stopped!

4. Case study

Symbolic function identification is often used as an illustrative example for evolutionary methods, especially genetic programming (Koza, 1992). Although simple regression problems are quite quickly solved by most GP implementations, more complicated or noisy problems remain a challenge. The presented case study focuses on the robustness of the evolved symbolic functions.

AEA can easily be used to evolve genetic programs. The standard GP representation of an individual – a tree-like structure – is convenient also for AEA. The tree consist of a number of nodes, each node representing one instruction to be executed. Size of the individual corresponds to the number of nodes in the tree.

The *success predicate* introduced by Koza requires perfect knowledge whether the solution is correct. The symbolic regression with *noisy* data set does not have a perfect fitness function nor perfect termination criterion. In order to determine the probability of satisfying the problem's success predicate, Koza measured the number of processed individuals. Here, we'll use the number of "function executions" as a measure of processing done by both algorithms.

4.1 Symbolic regression

Objective of this study is the discovery of a symbolic expression that satisfies a set of data points. Target function is the well known $t(x) = x^4 + x^3 + x^2 + x$ (Koza, 1992). In a perfect

learning environment GP excels in finding an exact solution because the data set includes no noise and an obvious fitness function is available.

To test the robustness of standard GP (SGP) and AEA-GP produced solutions, three data sets (L_0, L_1 and L_2) were created. Each data set included 41 points {(x_i,y_i) with $x_0 = -1$, $x_{i+1} = x_i + 0.05$}. In L_0 the dependent values y_i equaled $t(x_i)$. In L_1 and L_2, however, Gaussian noise $N(\mu,\sigma)$ was added. For L_1 and L_2 the dependent variables were $y_i = t(x_i) + N(0,0.02)$ and $y_i = t(x_i) + N(0,0.5)$, respectively. Figure 5 shows the data points of all three learning data sets.

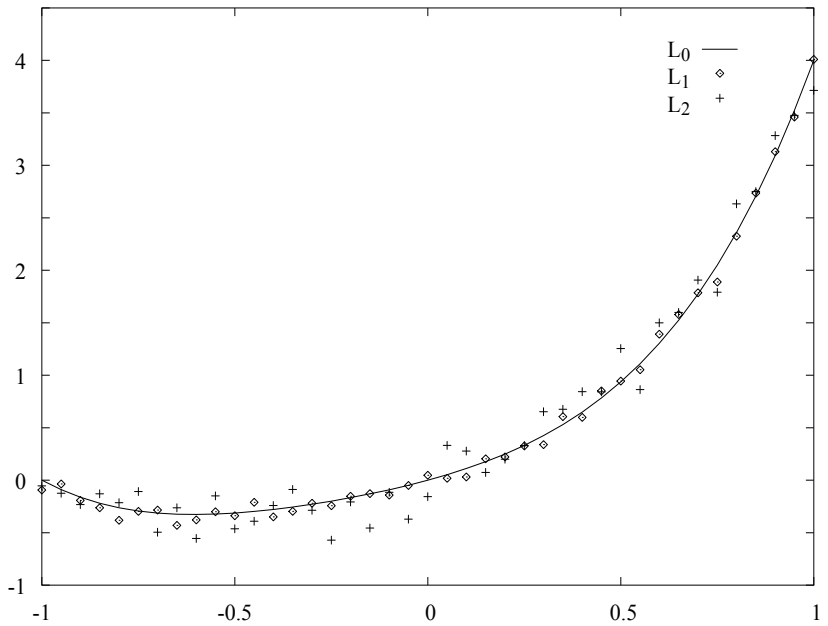


Fig. 5. The three data sets L_0, L_1 and L_2.

4.1.1 Standard GP

All SGP runs in this section were done by the OpenBeagle library (Gagne & Parizeau, 2006). When not explicitly stated otherwise, the default settings were population size 50, tournament selection 2, mutations 0.05 (including shrink-mutation), swap-sub-tree mutation 0.5, standard random-node mutation 0.05, crossover 0.9 and crossover point inside a tree 0.7. OpenBeagle used an adjusted fitness measure based on the accumulated error. The fitness value F of a symbolic function f on data L was calculated by

$$F(f,L) = \frac{1}{1+\sqrt{\sum \frac{(f(x_i) - L[x_i])^2}{|L|}}} \tag{5}$$

Another important criteria for a solution is size. In general, however, it is impossible to say whether a function comprising, for example 21 nodes, is always superior to the one comprising 25 nodes. Another problem are the introns. The effort to count function's introns would require additional processing and would slow down the evolution considerably. Because of this we chose to use the simple adjusted fitness measure only.

The described fitness measure needs to execute the function f exactly $|L|$-times, in our case 41-times. With the non-corrupted data set (L_0), the target function $t(x)$ had a fitness of 1. With noise the situation changed. Properties of the three data sets (L_0, L_1 and L_2) and the fitness values of target function $t(x)$ are shown in table 1.

data set	y	$F(y, L_i)$
L_0	$t(x)$	1
L_1	$t(x) + N(0, 0.05)$	0.9588
L_2	$t(x) + N(0, 0.2)$	0.8470

Table 1. Properties of the three data sets.

Both, OpenBeagle and AEA-GP, were equipped with the same function (FS) and terminal (TS) set:

```
FS = {+, -, *, \%, SIN, COS, EXP, RLOG},
TS = {x}
```

where '%' and 'RLOG' are protected division and protected logarithm, respectively. In order to choose the final solution from the evolved population, and to allow for a comparison with OpenBeagle, AEA-GP was equipped with the fitness measure (5).

4.1.2 OpenBeagle results

OpenBeagle offers a large number of available run-time options. The tests were performed using four different population sizes (P=50, 100, 200, 500). The evolution was interrupted if fitness hit the maximum or if the total number of fitness evaluations exceeded 1 million. Each configuration setup was used for 10 independent runs and the best-of-run individuals were recorded. Table 2 shows statistics for the hall-of-fame individuals of all 10 runs in each

data set	min	max	$mean \pm \sigma$	$t(x)$
P=50				
L_0	0.9990	1.0000	0.9999 ± 0.0003	9/10
L_1	0.9704	0.9797	0.9744 ± 0.0034	0/10
L_2	0.8732	0.9295	0.9110 ± 0.0173	0/10
P=100				
L_0	0.9977	1.0000	0.9998 ± 0.0007	9/10
L_1	0.9540	0.9802	0.9706 ± 0.0081	0/10
L_2	0.9060	0.9513	0.9228 ± 0.0146	0/10
P=200				
L_0	0.9959	1.0000	0.9993 ± 0.0015	8/10
L_1	0.9628	0.9801	0.9742 ± 0.0053	0/10
L_2	0.9024	0.9441	0.9233 ± 0.0141	0/10
P=500				
L_0	1.0000	1.0000	1.0000 ± 0.0000	10/10
L_1	0.9597	0.9767	0.9707 ± 0.0054	0/10
L_2	0.8937	0.9310	0.9146 ± 0.0124	0/10

Table 2. OpenBeagle-GP scores for best-of-run individual's fitness values for respective data sets using different populations sizes, averaged over 10 independent runs.

category. It shows results for respective data sets (column 1), the minimum achieved fitness (column 2), maximum fitness (column 3), mean fitness with standard deviation (column 4) and count of runs producing the target $t(x)$ (column 5).
The clean learning data set L_0 was not problematic – OpenBeagle found the target $t(x)$ in 36 out of 40 runs. If the population size was set to 500 it even scored 10/10!
Due to the noise in L_1 and L_2, the fitness function was unsuccessful in recognizing the target $t(x)$ and headed straight towards functions that were over-fitted to the learning data. In 80 runs it never recognized $t(x)$ as the final solution and only once evolved a solution with a fitness score below $F(t(x))$. For both, L_1 and L_2, GP almost always encountered the function $t(x)$ *during* the run, but discarded it and proceeded towards greater fitness. This was because the chosen (or any other) fitness function could not compensate for the noise in the data.

4.1.3 Autonomous EA

The interaction between the two opposing AEA-GP individuals was based on the comparison of the absolute error both candidate functions made on one random learning instance. AEA-GP was set to terminate artificially if the number of function executions reached 41 million executions (one fitness evaluation in OpenBeagle was a calculation of 41 function values, thus the SGP run executed at most 41*1M=41M functions). Of course, AEA GP also auto-terminated if any population lost all of its members ($P \to 0$). At the end, the population was inspected and, according to the SGP's fitness measure, "best" solutions were recorded.
Table 3 shows statistics for best-fitness individuals averaged over 10 independent runs for three different memory settings per respective data set. The minimum, maximum, mean and standard deviation of highest fitness values at the end of each AEA run are presented. Additionally, the average number of interactions $\overline{I}$ pro run is displayed. Column 5 shows the count of perfect solutions t(x) produced in auto terminated runs ($P \to 0$). Last column counts the number of runs producing target function t(x) regardless of the termination criteria.

data set	min	max	$mean \pm \sigma$	$\overline{I}$	$t_{P\to 0}/P \to 0$	$t(x)/n$
M = 50000, I_{max}=41M						
L_0	0.8171	1.0000	0.9353 ± 0.0840	≈ 8.8M	4/6	6/10
L_1	0.8195	0.9588	0.8907 ± 0.0597	20.5M	0	4/10
L_2	0.7819	0.8553	0.8081 ± 0.0223	20.5M	0	0/10
M = 250000, I_{max}=41M						
L_0	0.8355	1.0000	0.9550 ± 0.0728	≈ 16.2M	3/3	7/10
L_1	0.8562	0.9588	0.9434 ± 0.0324	20.5M	0	7/10
L_2	0.7785	0.8573	0.8215 ± 0.0302	20.5M	0	0/10
M = 500000, I_{max}=41M						
L_0	0.8476	1.0000	0.9706 ± 0.0620	≈ 15.0M	4/4	8/10
L_1	0.9077	0.9600	0.9403 ± 0.0215	20.5M	0	4/10
L_2	0.7913	0.8676	0.8419 ± 0.0239	20.5M	0	0/10

Table 3. AEA statistics for best-of-run solutions with respective memory settings M.

Like SGP, AEA-GP was also successful in finding the target function $t(x)$ in problem set L_0. The more memory was available, the better were the results. The first row of table 3 shows that AEA-GP produced target function $t(x)$ 6 times in 10 runs; in 4 out of 6 cases, the perfect solution $t(x)$ was found after the evolution auto-terminated; in two cases the $t(x)$ was present in

the population P after the time-out of 20.5M interactions (2 evaluations per interaction make 41M evaluations). Interestingly, 2 auto-terminated runs did not produce $t(x)$. This could be attributed to smaller populations; M, set too low, increased the probability of the type $S_1 - S_3$ scenario (Fig. 3). Average AEA run on L_0 with M = 50000 nodes took 8.8M interactions.

When describing L_1 and L_2 with 100% accuracy, no straightforward function exists. The termination criterion $P \to 0$ was even less likely to be satisfied (as was the case with L_0) as the noise disturbed each partially winning function. Consequently, AEA-GP always terminated only after 20.5M interactions and most[6] fit individual was the pronounced solution of the run. With L_1, AEA-GP was able to find target function $t(x)$ in 15 out of 30 runs. SGP, on the contrary, almost always ended up with an excessively over-fitted solution; only in one run out of 40 did it evolve a population with a maximal fitness *lower* than 0.9588. AEA-GP managed just the opposite: it evolved a solution with a fitness *higher* than 0.9588 in one run only.

With L_2, AEA-GP did not find target $t(x)$ in any of the 30 runs, but neither did SGP. SGP encountered several $t(x)$ quite early in the run but then discarded them in favor of other excessively over-fitted solutions. AEA, on the contrary, evolved towards $t(x)$ but failed to produce the desired target at the termination time[7].

Interestingly, with L_2, SGP's terminal population included the $t(x)$ once, yet SGP failed to *recognize* it. AEA-GP also saw $t(x)$ during the evolution. Due to the noise in L_2, however, the target vanished and was not present at termination time. If present, it would have been mostly recognized because the best individuals' mean fitness (0.8419) was always lower than $F(t(x))$ = 0.8470. SGP's achieved minimum fitness was always over-fitted well above that (0.8732).

4.2 Remarks

When comparing AEA with other evolutionary computing techniques, e.g. GP, special attention should be paid to the interpretation of the inherent time-line. GP terminates immediately upon encountering the first solution with the perfect fitness score (e.g. f = 1). Autonomous EA, on the contrary, auto-terminates only when P gets exhausted – this may have been long after the first 100% solution (f = 1) is found. It can be said that AEA is more time consuming than EA even though the AEA's interaction operator is mostly much faster than the full fitness calculation. On the other hand, the presented termination criterion is more problem independent than the common generational and/or success-oriented predicates, especially when without a perfect measure of quality – the case of most real-world problems.

Therefore, the traditional fitness-based EAs should be preferred over AEA if the search space is free of noise and if the learning set includes *all* possible instances and if it remains of manageable size (e.g. the n-multiplexer problem etc.). Namely, EAs converge faster if the solution can be described perfectly by a fitness function. In such cases their tendency to over-fit the training data is not problematic.

5. Conclusions

The presented AutonomousEA exploits and simplifies existing EA philosophy. It is based on a simulated interaction between two populations competing in a tournament-like manner.

[6] Again, fitness F is not optimal, but is most convenient.

[7] In some of the unofficial runs set to terminate at 40M interactions, AEA eventually produced $t(x)$ even in L_2.

The main loop guarantees each individual a momentarily action (life!) in a quasi-parallel style. The core algorithm continuously creates interactions of two random individuals from two opposing populations and takes care, that every individual gets its turn as soon as possible. The dynamic sizing of both populations is implicitly governed by the self-regulating principles (as in the predator-prey system), which determine the number of offspring and amount of mutations. The population size is therefore a direct consequence of individuals' ability to survive. This is how the AutonomousEA creates a link between the immeasurable qualitative and measurable quantitative properties.
Both initial random populations always include incompetent individuals, which are terminated during the evolution yet are necessary to shape the content and size of the population. Most of the processing time is used to create a population with precisely the right density of quality solutions. Higher density increases the probability of successful crossover and the creation of even better offspring, which in turn eliminates all individuals from the competitive environment. Empty environment is the termination criterion and signals a successful completion of the evolutionary search.
The AEA is based on the co-evolution of two populations of the individuals of the same type. The smaller (endangered) population evolves faster because all of its members are always active thus they have additional breeding chances compared to the larger population.
The concept of autonomous evolutionary algorithm needs only three run-time parameters:

- Initialization setting is a parameter needed in the initialization phase of the algorithm. It can be used to tailor the first evolvable objects (for example the GP tree initial depth setting).
- Memory space limits the population's size. The larger the value, the larger the two populations. This setting should be set high to fully exploit the available hardware, because large population sizes do not result in over-fitting problems as in standard EAs.
- Processing time is the artificial termination criterion because certain results must be delivered in due time.

AutonomousEA is very simple to *run* – the set-up phase is very similar to that of an EA yet for the run-phase only the memory space must be specified, all other details are self regulated. Traditional EAs, on the contrary, require much effort to determine just-the-right values for numerous parameters of the run phase.
Natural evolution is not under pressure to discover or optimize something. Rather, it goes different ways and something always pops out. The engineer, on the contrary, must hold artificial evolution in one direction. Use of fitness to specify this direction is problematic unless we're unable to create perfect fitness, because evolution will find a solution with best fitness but of small value.
The last section documented the AEA's performance in evolving genetic programs for the noisy symbolic regression problem. Comparison with standard GP gave an insight into the power of the AEA's principles regarding generalization capabilities of the produced solutions. Main problem of EAs was their tendency to over-fit the training data – the longer they were allowed to run or the larger the population size, the larger the discrepancy. AEA, on the contrary, kept close to the global generalization level. It found better solutions when more processing power (time) or memory was available.

5.1 Future work

Interesting sub-project was to use the adjusted fitness F instead of atomic tasks to decide the winner of an interaction. Although not in line with the AEA philosophy, it proved beneficial

for the symbreg example. On more difficult problems, however, it showed typical fitness-related problems (over-fitting. . .).
Also, different types of populations could compete under the AEA – for example a population of solutions could compete against a population of problems, what would allow for very elegant interaction implementation. This option, however, is problematic for many problem domains as it leads to premature convergence problem with instant AEA auto-termination. Fact is that both population must evolve in parallel. We cannot create a random initial population and expect it to solve difficult problems in the first try. A mechanism that would allow for "evolution" of problems is needed.

5.2 The code

Autonomous evolutionary algorithm relies on the best-ideas-are-simple philosophy – better EAs are more simple, not more complex. The AEA library and example projects in C++ are available from the author per email request (matej.sprogar@uni-mb.si).

6. References

Angeline, P. (1995). Adaptive and self-adaptive evolutionary computations, *in* M. Palaniswami & Y. Attikiouzel (eds), *Computational Intelligence: A Dynamic Systems Perspective*, IEEE Press, pp. 152–163.

Angeline, P. & Pollack, J. (1994). Competitive environments evolve better solutions for complex tasks, *Proceedings of the 5th International Conference on Genetic Algorithms (GA-93)*, pp. 264–270.

Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice*, Oxford University Press.

Banzhaf, W., Nordin, P., Keller, R. & Francone, F. (1998). *Genetic Programming – An Introduction*, Morgan Kaufmann, San Francisco.

Blanchard, P., Devaney, R. & Hall, G. (1997). *Differential Equations*, Brooks/Cole Publishing Company.

Bremermann, H. (1962). Optimization through evolution and recombination., *Self–Organizing Systems* pp. 93–106.

Dictionary.com (2006). fitness, *Based on the Random House Unabridged Dictionary* . URL: *http://dictionary.reference.com/browse/fitness*

Eiben, A., Hinterding, R. & Michalewicz, Z. (1999). Parameter control in evolutionary algorithms, *IEEE–EC* 3(2): 124.

Friedberg, R. M. (1958). A learning machine, part I, *IBMJ. of Research and Development* 2: 2–13.

Friedberg, R.M., Dunham, B. & North, J. (1959). A learning machine, part II, *IBMJ. of Research and Development* 3: 282–287.

Gagne, C. & Parizeau, M. (2006). Genericity in evolutionary computation software tools: Principles and case-study, *International Journal on Artificial Intelligence Tools* 15(2): 173–194. URL: *http://www.worldscinet.com/109/15/1502/S021821300600262X.html*

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison– Wesley.

Henle, K. (1991). Some reflections on evolutionary theories, with a classification of fitness, *Acta Biotheoretica* 39(2): 91–106.

Hillis, D. (1990). Co–evolving parasites improve simulated evolution as an optimization procedure, *Physica D* 42: 228–234.

Holland, J. (1975). *Adaptation in natural and artificial systems*, The University of Michigan Press, Ann Arbor, MI.

Holland, J. (1995). *Hidden Order – How Adaptation Builds Complexity*, Addison–Wesley Publishing Company.

Juille, H. & Pollack, J. (2000). Coevolutionary learning and the design of complex systems, *Advances in Complex Systems* 2(4): 371–394.

Koza, J. (1992). *Genetic Programming: On the Programming of Computers by Natural Selection*, MIT Press, Cambridge,MA.

Mandelbrot, B. & Hudson, R. (2004). *The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin and Reward*, Basic Books.

Pagie, L. & Hogeweg, P. (2000). Information integration and red queen dynamics in coevolutionary optimization, *Proceedings CEC 2000*, pp. 1260–1267.

Pagie, L. & Mitchell, M. (2002). A comparison of evolutionary and coevolutionary search, *International Journal of Computational Intelligence and Applications* 2(1): 53–69.

Paredis, J. (1997). Coevolving cellular automata: be aware of the red queen!, *in* T. Bäck (ed.), *Proceedings of the 7th Int. Conference on Genetic Algorithms (ICGA 97)*, pp. 393–400.

Price, K. & Storn, R. (1997). Differential evolution: A simple evolution strategy for fast optimization, *Dr. Dobbs Journal of Software Tools* 22(4): 18–24.

Sareni, B. & Krähenbühl, L. (1998). Fitness sharing and niching methods revisited, *IEEE Transaction on Evolutionary Computation* 2(3): 97–106.

Stanovich, K. E. (2003). The fundamental computational biases of human cognition: Heuristics that (sometimes) impair reasoning and decision making, *in* J. E. Davidson & R. J. Sternberg (eds), *The psychology of problem solving*, Cambridge University Press, pp. 291–342.

Toffoli, T. (2000). What you always wanted to know about genetic algorithms but were afraid to hear. URL: *http://www.citebase.org/abstract?id=oai:arXiv.org:nlin/0007013*

Wolpert, D. &Macready,W. (1997). No free lunch theorems for optimization, *IEEE Transactions on Evolutionary Computation* 1(1): 67–82.

17

Development and Evaluation of the Spoken Dialogue System Based on the W3C Recommendations

Stanislav Ondáš and Jozef Juhár
Technical University of Kosice, Faculty of Electrical Engineering and Informatics
Slovakia

1. Introduction

Due to progress in technology of speech recognition and understanding, the Spoken dialogue systems (SDS) have started to emerge as a practical alternative for a conversational computer interface. They are more effective than Interactive Voice Response (IVR) systems since they allow a more free and natural interaction. The Spoken dialogue systems are designed for providing automatic dialogue-based voice services accessible through telephone. Such systems consist of a number of components that need to work together for the system to function successfully (McTear, 2005). The basic architecture of the SDS consists of (more or less indispensable) modules – dialogue manager, language understanding, speech recognition, access device interface, language generation and text to speech synthesis. Easiness of implementation and rapid development of voice services led to the standardization effort. The World Wide Web Consortium plays an important role in this area.

The book chapter proposed will be focused on design, development and evaluation of the Spoken dialogue systems based on the W3C Recommendations. The World Wide Web Consortium is an international community that develops standards to ensure the long-term growth of the Web (W3C, 2010). One of their workgroups, Voice Browser Working Group, deals with preparing standards for voice-enabled technologies. The main idea is to build "Voice browser" enabling access to the information by voice, similarly as in the case of web browser. Comparison of definitions of the Spoken dialogue system and Voice browser lead to the conclusion, that both systems are very similar or de facto identical. A group of XML-based languages (SIF - Speech Interface Framework) was defined by Voice Browser Working Group to enable speech communication between user and computer. The W3C SIF recommendations became the industry standards in voice-enabled technology domain during the last decade. Languages in SIF also define the interfaces between fundamental subsystems of Spoken dialogue system and thus determine the basic structure of such system. The main languages in the framework are VoiceXML, SRGS and SSML that enable composing dialogues, speech grammars and instructions for text-to-speech systems. The CCXML serves for handling I/O (telephony) devices. The SISR specification defines the semantic tags for speech grammars to enable extracting of the meaning of user's input. The meaning can be represented in the EMMA language, which was prepared by the W3C

Multimodal Interaction Working Group (MIWG, 2010). The PLS defines tags for composing pronunciation lexicons for ASR as well as TTS systems.
Evaluation of the Spoken dialogue systems and their services is also very important in the system's life-cycle. During the test phase it may bring the information about the performance of the system, in the phase of pilot running enable to observe the impact of changes, which were done and in the phase of public running can provide an estimation of the quality perceived by the users.
At the beginning of the article we will provide description of these languages and some other important technologies and then we will outline the architecture of SDS, which is based on W3C recommendations. Then, in section three, the research and development of the Slovak spoken dialogue system (SDS) will be described, which adopts several W3C languages. The architecture and components of the system will be introduced. The last part of the article will be focused on objective as well as subjective evaluation methods. Both methods bring different information about the system and services being provided. The objective method based on collecting of interaction parameters will be presented. For that purpose the evaluation server was integrated into the Slovak SDS. It is described in subsection 3.2.6. Also a subjective evaluation based on filling in the questionnaire was carried out and will be described in the section four.

2. Description of the W3C Speech Interface Framework

The World Wide Web Consortium (the W3C) is an international community developing the standards ensuring the long-term growth of the Web (W3C, 2010). The Consortium consists of working groups associated with the research area. One of them is the Voice Browser Working Group (VBWG) focused on development of standards for Voice Browsers. Voice browsers allow people to access the Web using speech synthesis, pre-recorded audio, and speech recognition through their phone device. The Voice Browser Working Group was first established on March 26, 1999 (VBA, 2010), to develop specifications for these devices. The W3C Speech Interface Framework (SIF) is a suite of markup specifications aimed at realizing this goal. Languages in that group fulfil the idea of portability and rapid development of voice services. The framework actually consists of VoiceXML, SRGS, SSML, PLS, SISR, CCXML and SCXML specifications. Following subsections provide their short description.

2.1 Voice eXtensible Markup Language

The VoiceXML (Voice eXtensible Markup Language) is a markup language designed for composing the voice applications. In 1999 four companies AT&T, IBM, Lucent and Motorola established the VoiceXML forum (VXMLForum, 2010) for designing a language, which would increase the development of voice applications. The first version of the language was introduced in august 1999. The first official version of the VoiceXML language (VoiceXML 1.0), prepared by VoiceXML forum, was presented in March 2000. After that the W3C adopted the responsibility for VoiceXML language, and it had started working on the next version of VoiceXML.
Whereas the VoiceXML 1.0 language specification implied, besides tags (markups) for dialogue description, also tags for call management, speech grammars and speech synthesis, the second version of the language (VoiceXML 2.0) was focused only on dialog description. Markups for call management, speech grammars and speech synthesis control were adopted as the background for CCXML, SRGS and SSML languages. VoiceXML 2.0 was released as

the W3C recommendation in March 2004. This recommendation became the industry standard in area of voice services.
In June 2007 the VoiceXML 2.1 was introduced, which attaches a tiny set of additional features to the second version of the language. Then working on the new specification (VoiceXML 3.0) has started, with the concept of three layers - dialog, flow and management. Work on the third version of the language is still in progress.

2.2 Speech recognition grammar specification

As mentioned above, a tiny set of markups used in VoiceXML 1.0 language has created a base of Speech Recognition Grammar Specification (SRGS). SRGS specification brings a language, which enables arranging context-free grammar for speech or DTMF input. Grammar can be specified in either XML or an equivalent augmented BNF (ABNF) syntax. Work on this language has been started in 1999 and it became the recommendation in March 2004 (SRGS 1.0).
The main advantage of the SRGS is well readable form both for designers and computers. It enables composing possible language structures, that are expected from user in actual state of interaction (dialog). Creation of such structures helps the speech recognition system to be more accurate and faster.
SRGS specification can describe (handle) also speech input in a form of utterances in natural language, but it does not support stochastic language models (N-grams) directly. The N-gram specification serves for that purpose, but it has never been published as the W3C recommendation and its preparation did not continue.
The power of SRGS specification is in cooperation with the next W3C specification – Semantic Interpretation for Speech Recognition (SISR).

2.3 Semantic interpretation for speech recognition

The semantic interpretation specification describes annotations to grammar rules for extracting the semantic results from recognition. This provides markups and attributes, which can be included in to context-free grammar and thus some semantic information can be extracted by interpretation of these markups. De facto, it does not really "understand", but it is the acceptable approach to the interpretation of spoken language. Such approach can be used also with the input utterances in natural language. In this case, the system can be viewed like keyword-spotting system. It enables capturing keywords in a natural language utterance and assigning them some semantic value and creating pairs of keywords and their semantic values. This concept is very powerful in domain-specific voice services, but almost unusable in communication with conversational agents.
Work on this specification had started in April 2003 and in April 2007 it became the W3C recommendation.

2.4 Pronunciation Lexicon specification

Pronunciation Lexicons describe phonetic information for use in speech recognition and synthesis. The requirements were first published on March 12, 2001, and updated on October 29, 2004. The pronunciation lexicon is designed to enable developers providing supplemental information on pronunciation for items as are place names, proper names and abbreviations. The W3C Recommendation was published in October 2008. Such lexicon can be used both by automatic speech recognition systems and text-to-speech systems.

2.5 Speech Synthesis Markup Language

As in the case of SRGS specification, designing the Speech Synthesis Markup Language (SSML) has started in year 1999. This process led to the first recommendation of the language (SSML 1.0) in September 2004. Work on this language is still not finished. The SSML 1.1 specification provides a tiny set of additional features to make this language more usable. The speech synthesis specification (SSML) defines a markup language for prompting users via a combination of pre-recorded speech, synthetic speech and music. It provides uniform API between voice platforms and Text-to-Speech engines and enables changing voice characteristics, like gender, speed, volume, etc.

2.6 Call Control eXtensible Markup Language

The W3C is designing the Call Control eXtensible Markup Language (CCXML) to enable fine-grained control of speech (signal processing) resources and telephony resources in a VoiceXML telephony platform. CCXML is designed to manage resources in a platform on the telecommunication network edge. It can handle actions like call screening, call waiting/answering and call transfer.

Requirements for that language were prepared in April 2001 and now the language has status "recommendation candidate" (April 2010). This specification brings very important unification into the call traffic handling, because of large range of telephony hardware producers. It releases voice services designers from concerning about hardware-specific application interface and it gives them the high-level interface by the CCXML language.

2.7 State Chart eXtensible Markup Language

A State Chart XML or the *State Machine Notation for Control Abstraction* is the last part of the Speech Interface Framework. SCXML is a candidate for being the control language within VoiceXML 3.0, the future version of CCXML, and the multimodal authoring language. Its development started in July 2005 and currently the seventh working draft was published. This new specification is connected to the new idea of *data-flow-management* framework. The main idea is separation of these three layers, because of higher transparency.

2.8 Extensible MultiModal Annotation markup language

The Extensible MultiModal Annotation markup language (EMMA) is a markup language intended for use by systems that provide semantic interpretations for variety of inputs, including but not necessarily limited to, speech, natural language text, GUI and ink input (MIWG, 2010). It provides a group of tags for describing semantic of such inputs. The language is developed by the *W3C Multimodal Interaction Working Group* (MIWG, 2010), which aims at developing specifications to enable accessing the Web using multimodal interaction. The first working draft for this specification was published in August 2003 and it became the recommendation in February 2009.

2.9 The W3C-based architecture of the voice browser

The W3C Speech Interface Framework languages have their main employment in voice browsers as well as in spoken dialogue systems. The languages from SIF determine the key ideas about cooperation between voice browser components; de facto they determine the architecture of such system.

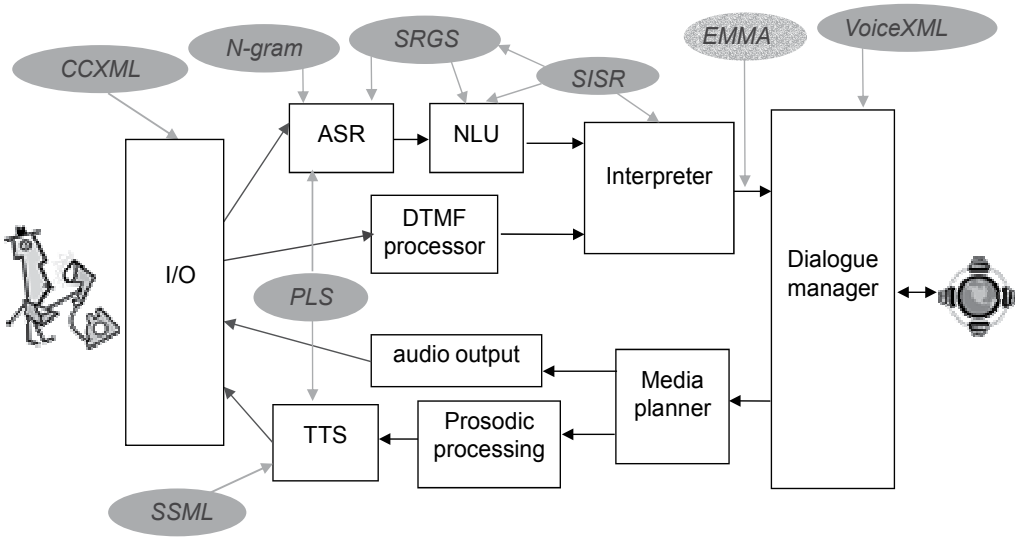


Fig. 1. General structure of the voice browser architecture.

At the end of year 1999 Voice Browser Working Group published working draft of document "Model Architecture for Voice Browser Systems" (MAVBS, 1999). Authors of the document allege that they only wanted to illustrate one of possible solutions of the Voice browser architecture and that the other types of architecture can be adopted for implementation of Speech Interface Framework languages. Interfaces between voice browser's components were not specified directly. The languages within the SIF determine the way of communication between them, what is the main advantage of the Speech Interface Framework. The model architecture of Voice browser from document mentioned above, redrawn into well arranged form in (Delgado, 2005), with stand-alone I/O component is displayed on Fig. 1.

The shadow ellipses represent SIF languages, which should be supported by voice browser components. The main components of the browser are dialogue manager, automatic speech recognition system (ASR), text-to-speech system (TTS) and Input/Output component. The NLU component is responsible for extracting the meaning from user's input. DTMF processor enables processing the DTMF input and both types of input (speech/DTMF) are finally processed in the Interpreter. On the other side, there are Media planner component and audio output block. A block of the prosodic processing is often the part of TTS system.

3. The W3C based Slovak spoken dialogue system

The Slovak spoken dialogue system has been developed in period from July 2003 till June 2006 and was supported by the National program for R&D "Building of the information society". The main goal of the project was the research and development of the SDS for information retrieval using voice interaction between humans and computers. The SDS had to enable multi-user interaction in Slovak language through telecommunication networks and to find information distributed in computer data networks such as the Internet. The SDS is also a tool for continuous research in the area of spoken language technologies in Slovakia (Juhár et al., 2006).

The choice of the solution sourced from contemporary free resources, state-of-the-art in the topic and the experiences of the partners involved in the project. Portability and easiness of compiling new services were considered as important factors. The final solution is based on the DARPA Communicator architecture with the central hub process, a software router developed by the Spoken Language Systems group at MIT, subsequently released as an open source package in collaboration with the MITRE Corporation, and now available on SourceForge (Polifroni & Seneff, 2000). The architecture of the system was designed to be compatible with the W3C Speech Interface Framework. The proposed system consists of a Galaxy hub and six modules (servers).
Since 2006 the Slovak SDS is being improved continually at Technical University of Košice with collaboration of Slovak Academy of Science in Bratislava.

3.1 System architecture

The architecture of the developed system uses a 'hub-and-spoke' architecture: each module seeks services from and provides services to the other modules by communicating with them through a central software router - the Galaxy hub. Mentioned system (Fig.1) consists of a hub and six system modules: telephony module, automatic speech recognition (ASR) module, text-to-speech (TTS) module, back/end module (Information server), module of dialogue management and the evaluation module. The relationships between the dialogue manager, the Galaxy hub, and the other system modules are represented schematically in Fig. 1.

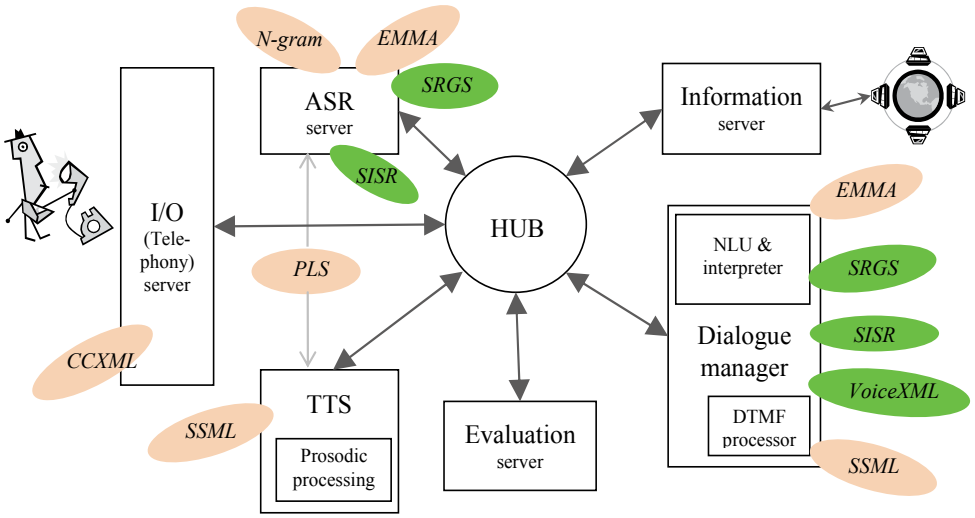


Fig. 2. Architecture of the Galaxy/W3C based Slovak spoken dialogue system

The white rectangles depict the system modules. Green ellipses show the W3C languages, which are supported by Slovak SDS. The orange ellipses show the W3C languages, which can be implemented for full support of the W3C SIF. As we can see on the Fig. 2 some specifications must be supported by more than one system module. For example the SRGS specification must be supported by ASR server as well as Dialogue manager. Dialogue manager needs to generate SRGS grammars and needs to analyze them. The ASR server also needs to analyze SRGS grammar and use it in the recognition process.

As was said above the principle of communication in proposed architecture is the exchange of messages, which are requests on specific functionalities or replies to those requests. The key functionality of messages is to establish the connection, to create and transfer requirements on services of other servers, to transfer replies to those requirements, to inform other servers of the system and so on. In the SDS based on the W3C SIF languages, messages are also the way of transporting code of the SIF languages (part of codes, files, addresses), that hold all information, which is required for communication between servers. From that point of view, messages are only transporters of communication.

3.2 System modules

Proposed solution of the SDS system poses common requirements on the module's structure. Each module should have some sub-modules, specifically Galaxy interface, XML Parser and Resource Manager. The Galaxy interface is responsible for communication between module and hub process. It defines dispatch functions (functions, which are provided by the module to the system) and it handles incoming and outgoing messages. Almost all modules need to analyze XML-based documents, because the W3C languages are XML-based. Therefore an XML Parser is the indispensable sub-module. The last one is the Resource manager sub-module, which handles all resources (source files, audio files, models, etc.).

3.2.1 Telephony (I/O) server

A telephony module connects the whole system to telecommunication network. It opens and closes telephone calls and transmits speech data to/from the ASR/TTS modules through the broker channel. The server supports telephone hardware - Dialogic D120/41JCT-LSEuro voice board (Juhár et al., 2006), which creates an interface to PSTN and GSM network (through hardware GSM gateways). Nowadays VoIP (Voice over IP) technology is becoming widely used in telecommunication. The key role is played by the Session Initialization Protocol (SIP), developed by IETF, which is a text-based protocol, similar to HTTP and SMTP, for initiating interactive communication sessions between users (Rosenberg, 2006). The Slovak SDS is connected also to VoIP network by integrating the Open Source library PJSIP in the telephony module (Pleva et al., 2008). The server is ready for integrating of CCXML language, which should be responsible for the interaction between SDS and the telephony module and for the management within the module.

3.2.2 Automatic speech recognition server

The automatic speech recognition (ASR) server performs the conversion of incoming speech to the corresponding text. The ATK based speech recognition engine was adopted for our SDS. The ATK is an Application Toolkit for HTK, which is freely available for non-commercial research (Young, 2004). The Audio input module of the basic recognition system was substituted by the new one, with the interface attached to the Galaxy broker channel. Also there were functions added for creating a group of language sources (grammars and dictionaries), for converting the grammar format from SRGS to HTK-compatible format and backwards, for indicating of VoiceXML *noinput* and *nomatch* events etc. The ASR server in the Slovak SDS supports SRGS and SISR specifications.

Context dependent HMM acoustic models trained on SpeechDat-Sk (Pollak et al., 2000) and MobilDat-Sk (Rusko et al., 2006a) speech databases were used for recognition of the Slovak

language. Context dependent (triphone) acoustic models were trained in a training procedure compatible with "refrec" (Lihan et al., 2005).

3.2.3 Text-to-speech server

The text-to-speech (TTS) synthesis server converts outgoing information having the text form into the acoustic form. A concatenative text-to-speech synthesis engine for Slovak language was developed by Slovak Academy of Science. Diphones were selected as a good candidate at this type of synthesis for Slovak language. Two unit selection algorithms that create final sequences of diphones were prepared. Both are based on minimizing the number of artificial concatenations in a synthetic signal. The first algorithm propagates from the longest units to the shortest ones, first querying the corpus for the whole phrase, then for its sub-phrases and finally for the diphones. The second algorithm starts from the shortest units. It just queries the corpus for all of the diphones that are then put into a lattice of candidates. Finally Viterbi search is used to find the path through the lattice with minimum number of concatenations (Rusko et al., 2006b). The TTS server is prepared for supporting of the SSML language in the future.

3.2.4 Dialogue manager server

The main module in the Slovak SDS is the dialogue manager server (Ondáš & Juhár, 2005), which controls the interaction between system and user and it is responsible for generating requests to the system's actions like playing prompts, recognizing user's utterance, obtaining information from the Internet and etc. De facto it also manages other servers of the SDS. The dialogue manager module in the Slovak SDS fully supports VoiceXML 1.0 specification and significant part of VoiceXML 2.0 specification. Therefore the heart of the manager is an interpreter of VoiceXML mark-ups. The principle scheme of DM module is shown on Fig. 3. The core of the manager is the aforementioned VoiceXML interpreter together with ECMAScript engine and XML Parser, which directly relates to interpretation of VoiceXML scripts. Output generator is responsible for implementation of the Prompt selection algorithm and its output can have format of SSML document. The input processor together with the grammar manager constitutes the NLU subpart of the dialogue manager. Grammar manager integrates the Grammar activation algorithm as well as semantic

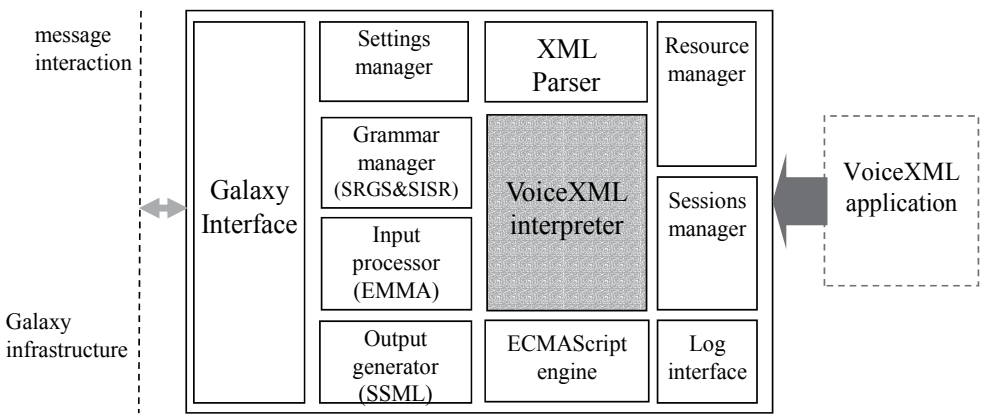


Fig. 3. The principle scheme of the Dialogue manager module

interpretation according SISR markups. The input processor generates requirements on acquiring user's input (DTMF or spoken utterance) and initializes processing of incoming inputs. Resource manager implements the Resource fetching algorithm. Settings manager is responsible for management of actual settings of the DM as well as SDS and for the realization of *property* markup of the VoiceXML language. Sessions manager creates, manages and destroys the user's session. Recording of information about interaction and internal state of DM is done by Log interface. The last part of the manager is the Galaxy interface, which connects it with remaining part of the system.
A dialog with the user starts after the message from the I/O server about the incoming call. Then DM initializes a new session and loads the VoiceXML script (application), interprets it and according to those it generates requirements on services (functions) of others system modules, in such manner, that the dialogue with the user is led.

3.2.5 Information server

Next part of the Slovak spoken dialogue system is the information server or backend server. It is capable of retrieving the information contained on the suitable web-pages according to the dialogue manager requests, extracting the requested data, analyzing them and if they are considered being valid, it returns the data to the DM. These tasks are performed by unique web-wrappers, which are the important parts of the server. The web-wrapper is responsible for the navigation through the web-server, data extraction from the web-pages and their mapping in a structured format (XML), convenient for further processing. In most cases the wrapper is specially designed for one source of data; thus for combining the data from different sources, several wrappers must be designed. Wrappers are designed to be as robust against changes in the web-pages structure as possible. To speed up the system (to eliminate the influence of long reaction times of the www-pages) and to assure drop-out resistance with simultaneous refresh of the information, automatic periodic download and caching of the web-pages content were introduced. The server is open to future applications by the possibility of creating web-wrappers for new services and adding them to the existing wrappers for Weather Forecast and Train Timetable services.

3.2.6 Evaluation server

The evaluation server tracks communication among servers and computes a set of interaction parameters. Möller in (Möller, 2005) defines interaction parameters as measures related to the system and the service which can be collected during the interaction. They enable quantifying parameters of the system and services and according to them concluding the quality of the evaluated system and services. Several interaction parameters can be obtained without human expert that leads to the automatic evaluation. This aspect is very important because of reducing costs spending on the evaluation during the system's lifecycle. Evaluation server in the Slovak SDS enables collecting the interaction parameters like dialogue duration, system prompt delay, user response delay, average number of noinput or nomatch events and so on. It produces several log files with those parameters for each channel and calculates overall as well as channel statistics.

3.3 Voice services provided by Slovak SDS

The Slovak SDS provides two pilot voice services – Weather forecast service for Slovakia and Timetable services (for Slovak Railways, Buses and City buses of Košice). The services

have been designed as a system-directed with the open structure based on subdialogues (Ondáš, 2007). The data are retrieved directly from the Internet. Provided services enable metacommunication in a form of a set of keywords like help, back, repeat, etc.
The Weather forecast service provides information about weather for thirty towns in Slovakia up to three days ahead.
The Timetable services allow finding out the connection between more than ten thousand points (stations or POIs) in Slovakia. After the welcome message and the service selection subdialog the user is prompted to provide information about start point, end point, date and time of the departure. In the designed dialog we have used explicit conditioned confirmation after a pair of input items. Using of the implicit confirmation require more sophisticated speech grammars, therefore we did not select this form of confirmation. Conditioned confirmation means that the input items collected from the user are confirmed only when their confidence level is lower than chosen threshold. After the confirmation of all provided information from the user, the system generates a query to retrieve data from internet, in which all items are sending to the web server. Then the obtained data are played to the user. Playing of the information is divided in to two layers. In the higher layer only the basic information is played. If the user wants to play detailed information, he can enter the lower layer, in which detailed information will be played.
The system is in public experimental running since January 2006. It runs in multi-user mode through PSTN and GSM telephony network and through VoIP (Skype&SIP). Until now more than 8000 calls were answered by the Slovak SDS.

4. Subjective evaluation of the Slovak SDS

Spoken dialogue systems (SDS) are nowadays widely used in several domains. This fact brings a need of evaluation, comparison and categorization of dialogue systems and their services. The quality of the interaction with a telephone-based speech service can be addressed from two separate points-of view. System developers are concerned about system/system's modules performance. From the user's point-of view, the perceived quality or overall opinion are the most important aspects (Möller, 2005). Using subjective measures is the only way of finding out user's opinion of the system. Objective measures, such as performance, do not have a direct attachment to the user satisfaction (Hartikainen et al., 2004). From that point of view, measurement of interaction parameters can bring only some information about system behaviour and properties. While extensive effort has been put to the definition and measurement of efficiency and effectiveness in user interactions with speech systems (ITU-T P.851, 2003), comparatively little emphasis has been put on measurement of subjective satisfaction (Hone & Graham, 2001). There are few projects, which are focused on this domain. The first one is the SASSI (Subjective Assesment of Speech-System Interface) methodology, which considers the "validity" and "reliability" aspects of the questionnaires as the most important. The questionnaires in this project are prepared in iterative design process, in which, at first, a pool of attitude statements is designed and in several iterations only a relevant set of statements is selected. An initial 50 item questionnaire was designed. Each attitude statement was rated according to seven points scale. Authors identify six factors or quality aspects: perceived system response accuracy, likeability, cognitive demand, annoyance, habitability and speed (Hone & Graham, 2001).
SERVQUAL method for subjective quality evaluation was adopted from the area of marketing applications and it is suitable also for subjective evaluation of dialogue systems.

Authors in (Hartikainen et al., 2004) view the spoken dialogue system as a service, which is provided to the users. This method is based on two principles: Service quality can be divided into dimensions, and measured as a difference of expectations and perceptions (Parasuraman, 1988). SERVQUAL method defines five service quality dimensions: tangibles, reliability, responsiveness, assurance and empathy. SERVQUAL method uses questionnaires by 22 items. Respondents assess how the reality meets their expectations.
The subjective evaluation methods providing information about the quality of telephone services are also described in the ITU Recommendation P.851 (ITU-T P.851, 2003). The evaluation methods described in this recommendation address different aspects of quality from a user's point of view, as are the usability of the service, the communication efficiency, task and service efficiency, user satisfaction, perceived speech input and output quality, the system's cooperability, etc (Möller, 2005). Described methods are based on laboratory experiments in which subjects interact with the spoken dialogue system in order to perform a pre-defined, realistic task. Then they fill a set of questionnaires, which reflected their opinion on assessed system and services. Recommendation contains also the set of questions (items) related to described aspects of quality and the examples of test scenarios.
Within years 2006 to 2008 we have proposed a new subjective evaluation method, which adopts methodology described in ITU P.851 Recommendation. There were several motivations for proposing a new (modified) method. The first, perceived quality of services provided by the Slovak dialogue system has never been adequately evaluated. Second motivation was that the methods introduced above cannot be simply used for evaluation of the Slovak SDS, because they do not take in to consideration usage of the W3C languages and Galaxy infrastructure. Our goal was also to propose the simple method with comparable and understandable result in a form of school grading system. The reasons for the selection of the ITU P.851 Rec. as a basis for the new method are summarized in (Ondáš et al., 2008).
The designed method is questionnaire-based and it is appointed to realization of subjective evaluation experiments for obtaining the user judgments and opinions of the spoken dialogue system and service quality. Within this method a set of questionnaires, test scenarios and rating scales are defined. Also there are defined both, new categorization of quality aspects and grades of quality for rating of these categories.

4.1 Evaluation questionnaires and methodology

The questionnaire form of evaluation was selected as the most suitable form of obtaining information about perceived quality of dialogue system and its services. From a set of questions/items, defined in ITU T. Rec P.851, three types of questionnaires were prepared:

- Questionnaire A - contains items related to user's background, their knowledge about domain and system. The items in this questionnaire were modified for testing the Slovak timetable information service. It contains 12 items.
- Questionnaire B - comprises 17 items related to individual interaction with the system.
- Questionnaire C - contains 14 closed and 3 open items related to the overall impression of the system and provided services.

Tab. 1. contains item numbers according to ITU T.Rec P.851, which was adopted for the designed questionnaires. For obtaining a complex view on quality aspects it is necessary to interact with the system more than once. Also the sufficient motivation on the side of the test subjects is required. Because of these facts, a set of test scenarios was designed. Each

user (test subject) makes two calls on the system's telephone number. First, the test subject fills in the questionnaire A. Then they make a call on the one of the telephone numbers of the Slovak SDS and in spoken interaction realizes given "common scenario". After the interaction they fill in the questionnaire B (B1), in which they assess the prior interaction. Then they make a second call and carry out the given "individual scenario". They assess the prior interaction in questionnaire B (B2). At last the test subject fills in the questionnaire C, in which they assess the both interactions and their overall impression of the system and its service.

Type of questionnaire	Item numbers
Type B	overall impression, 1, 2, 4-9, 11, 15 - 17, 19 - 22
Type C	1, 3, 4, 6 - 8, 10, 11, 13 - 21

Table 1. Items used for building of questionnaires B and C

4.2 Processing the results

The questionnaires processing consists of two operations - coding and categorization. The coding is a substitution of data by symbols, which will be used in statistical methods. The rating scale, which can hold values from 0 to 1, was designed for coding of items of the questionnaire. The "0" value represents the worst (the lowest) level of the quality (property). Conversely, the "1" value represents the best (the highest) level of the quality.

The items/questions in the questionnaires were coded with five types of scales:

- the growing 7-level scale
- the decreasing 7-level scale
- 5-level Likert scale
- backward 5-level Likert scale
- 5-level centered scale (the highest value in the middle)

There was designed categorization on six aspects of quality (categories). The first four categories are the same as in ITU T. recommendation. The last two categories were created for obtaining direct information about user's satisfaction and about usability of the system and its services. The designed categories are the following:

- *information obtained from the user; communication with the system; system's behaviour; dialog; user satisfaction; usability*

Each category is characterized by the set of questions from questionnaires B (1 and 2) and C. One question can be assigned to several categories. Then the categories are rated by one of the quality grades. There were six quality grades designed, from A to FX according to standard school grading system. For each category the quality grade is evaluated as an arithmetic mean of all item's responses for the given category (their score, assigned during coding) in percentage.

4.3 Evaluation experiments

Two evaluation experiments were performed with the basic setup of the Slovak spoken dialogue system. The first experiment was carried out with 26 test subjects (students). They made 52 interactions with the system. They filled in 104 questionnaires of A, B and C type. The calls were made through PSTN network in two acoustic environments - in the office (the

silent environment) and in the laboratory with twelve students (noisy environment). Evaluation was performed on Slovak railway timetable service. The second column in Table 2 contains the results of the experiment. Experiment is in more detail described in (Ondáš et al., 2008).

Category	Experiment 1	Experiment 2
Information obtained from the system	C (79.4%)	*C (79.5%)*
Communication with the system	D (65.9%)	C (75.1%)
Behaviour of the system	C (78.5%)	C (76.6%)
Dialog	C (71%)	D (67.9%)
User satisfaction	D (64.5%)	C (71.6%)
Usability	D (62.3%)	D (63.2%)

Table 2. Results of evaluation experiments

The second experiment was carried out with 24 test subjects. Evaluation methodology was reduced because of several facts. The most important fact was that all test subjects had some experience with evaluated SDS. Therefore, they did not fulfil questionnaire type A and they interacted with the system only once. Our experiences have shown that there is a high correlation between items in questionnaire B and C. Therefore, test subjects fulfilled only questionnaire type C and classification of the items was modified by mapping questions from that questionnaire. The proposed reduction in evaluation methodology led to absence of items for the first category. The value in this category was replaced by the interaction parameter *successfulness,* defined in (Ondáš & Juhár, 2007), which was obtained by expert evaluator. The experiment and evaluation was performed on City bus timetable service. The third column in the Table 2 contains results of the second experiment. The second experiment is in more detail described in (Ondáš et al., 2009).

5. Conclusion

The W3C Speech Interface Framework specifications have become industrial standards in domain of voice user interfaces, voice browsers as well as spoken dialogue systems. They enable rapid development of that systems and a large range of voice applications. The main advantages of the Speech Interface Framework consist of portability, uniformity of design process, easiness of designing the new systems and services and the possibility of automatic generation of such applications. A set of XML-based languages provides a good starting point also for integration with web applications.

Proposed article shows the possibility of combining free resources with up to date standards. Relatively old Galaxy infrastructure, used in the Slovak SDS, provides surprisingly great solution for integrating the W3C Speech Interface Framework languages, with the properties, that can lead to building the complete system for real environment. Of course, the success of such system relies on technologies employed in system servers (speech recognition, text-to-speech, etc).

In the future our work will be focused on supporting all SIF languages by our dialogue system as well as on enabling more natural interaction (advanced dialogue) between system

and user. Nowadays the interaction in the dialogue system has a form of filling in the questionnaire by voice. That means, the system ask the user for providing information, which determines its request and then it tries to obtain the answer from the web localities and deliver it to the user by speech synthesis. It is clear, that the nature of conversation is an interactive process rather than a structural product (Jokinen, 2009). The impossibility of asking the system by user is the main disadvantage of the VoiceXML language. The user mainly answers to the system questions. Of course, the answer of the user can be in a form of question, for example: "Could you help me?", but there do not exist the mechanism for real "user initiative". The VoiceXML language also allows so-called "interaction with mixed initiative", but this mode only enables user to provide several information in one utterance.
In our next work, we plan to focus on the solution for enabling interaction with user intitiative in the range of Speech Interface Framework.

6. Acknowledgements

The work presented in this paper was supported by Slovak Research and Development Agency under research projects APVV-0369-07 and VMSP-P-0004-09 and Ministry of Education of Slovak Republic under research projects VEGA-1/0065/10.

7. References

Delgado, López-Cózar R. & Araki M. (2005). Spoken, multilingual and multimodal dialogue systems: development and assessment. ISBN 0470021551, 9780470021552, John Wiley, 2005, Michigan university, 261 p., 2005

Hartikainen, M. & Salonen, E. & Turunen, M. (2004). Subjective evaluation of spoken dialogue systems using SERQUAL method. In ICSLP 2004. International conference on spoken language processing: proceedings. ICSLP, 2004

Hone, Kate S. & Graham, R. (2001). Subjective assessment of speech-system interface usability, In Eurospeech 2001, pp. 2083-2086, Scandinavia 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 3-7, 2001 ed. by Paul Dalsgaard, Borge Lindberg, Henrik Benner, and Zheng-hua Tan; ISCA Archive, 2001

ITU-T Rec. P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems, International Telecommunication Union, Geneva, 2003. http://www.itu.int/rec/T-REC-P.851-200311-I/en

Jokinen, K. (2009). Constructive dialogue modelling: speech interaction and rational agents, Wiley series in agent technology, ISBN 0470060263, 9780470060261, John Wiley and Sons, 2009, 160 p., 2009

Juhár, J. & Ondáš, S. & Čižmár, A. & Jarina, R. & Rusko, M. & Rozinaj, G. (2006). Development of Slovak GALAXY/VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet, Proceedings of Interspeech 2006, Pittsburg, USA, Sept. 17-21, 2006, paper 2056-Mon2FoP.10, 2006, Pittsburg

Juhár, J. et al.: Voice Operated Information System in Slovak, Computing and Informatics, Vol. 26, 2007, 577 - 603

Lihan, S. & Juhár, J. & Čižmár A. (2005). Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases, In Proc. Interspeech 2005, Lisabon, Portugal, September 2005, pp. 225 – 228, 2005

MAVBS, (1999). Model Architecture for Voice Browser Systems, W3C Working Draft: http://www.w3.org/TR/voice-architecture/, December 1999.

McTear, F., M. (2005). Spoken Dialogue Technology: Toward the Conversational User Interface, ISBN 1-85233-672-2, Springer-Verlag London Limited 2004, United States of America, 2004

MIT (2008). http://groups.csail.mit.edu/sls//technologies/galaxy.shtml, 2008

MIWG, (2010). Multimodal Interaction Working Group website: http://www.w3.org/2002/mmi/

Möller, S. (2005). Evaluating telephone-based interactive systems, In ASIDE-2005, Aalborg, Denmark, November 10-11. 2005, paper 42

Ondáš, S. & Juhár, J. (2005). Dialog manager based on the VoiceXML interpreter. In: Proc. of the DSP-MCOM 2005: The 6th international conference on Digital Signal Processing and Multimedia Communications, ISBN 80-8073-313-9, pp. 80-83, September 13-14, 2005, Košice, Slovak Republic, Technical university of Košice, Košice, 2005

Ondáš, S. (2007). Principles of voice services design for IRKR communicator, In: 7th PhD Student Conference and Scientific and Technical Competition of Students of Faculty of Electrical Engineering and Informatics Technical University of Košice: Proceeding from conference and competition, pp. 17-18, ISBN 978-80-8073-803-7, FEI TU Košice, May 2007, Košice

Ondáš, S. & Juhár, J. (2007). Automatic evaluation of Slovak spoken language dialogue system, In: ECMS 2007 & Doctoral School: 8th international workshop on Electronics, Control, Modelling, Measurement and Signals, pp. 91-96, ISBN 978-80-7372-218-0, Liberec, Czech Republic, May 21-23, 2007, Technical University of Liberec, 2007

Ondáš, S. & Juhár, J. & Čižmár, A. (2008). Evaluation of the Slovak spoken dialogue system based on ITU-T, In: Text, Speach and Dialogue: 11th International conference, TSD 2008, Brno, Czech Republic, September 8-12, 2008, Berlin, Springer-Verlag, ISBN 978-3-540-87390-7, ISSN 0302-9743, pp. 633-640, 2008

Ondáš, S. & Juhár, J. & Pleva, M. (2009). The city bus timtable voice service for Kosice, In: SPA 2009: signal processing: algorithms, architectures arrangements, and applications, ISBN 978-83-62065-00-4, pp. 154-158, Poznan, University of Technology, Poland, 2009

Parasuraman, A. & Zeithaml, V.A. & Berry, L.L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality, Journal of Retailing, 64, 1, 1988.

Pleva, M. et al. (2008). Concept of spoken dialogue system based on voice over IP telephony. In: RTT 2008: Research in Telecommunication Technology 2008, 9th International Conference, Bratislava, STU, 2008, ISBN 978-80-227-2939-0, pp 3, 2008

Polifroni, J. & Seneff, S. (2000). GALAXY-II as an Architecture for Spoken Dialogue Evaluation, Proceedings of Second International Conference on Language Resources and Evaluation (LREC), Greece, May 31-June 2, 2000, Athens

Pollak, P. et. al. (2000). SpeechDat(E) - Eastern European Telephone Speech Databases, In Proceedings LREC 2000 Satellite workshop XLDB - Very large Telephone Speech Databases, pp. 20-25, Athens, Greece, May 2000, Athens

Rosenberg, J. et al. (2006). SIP: Session Initialization Protocol. June 2002-2006, IETF, http://datatracker.ietf.org/doc/rfc3261/

Rusko, M. & Trnka, M. & Darjaa, S. (2006)a. MobilDat-SK - A Mobile Telephone Extension to the SpeechDat-E SK Telephone Speech Database in Slovak, In press: SPEECOM 2006, Sankt Peterburg, Russia, July 2006, Sankt Peterburg

Rusko, M. & Trnka, M. & Daržagín, S. (2006)b. Three Generations of Speech Synthesis Systems in Slovakia, In: Proc. of XI International Conference Speech and Computer, SPECOM 2006, Sankt Peterburg, Russia, 2006. ISBN 5-7452-0074-X, pp. 297-302.

VBA, (2010). Voice Browser Activity website: http://www.w3.org/Voice/

VXMLForum, (2010). VoiceXML Forum website: http://www.voicexml.org/

Walker, A. M. & Litman, J. D. & Kamm, A. C. & Abella, A. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents, in Proceedings of ACL/EACL 35th Annual Meeting of the Association for Computational Linguistics, San Francisco: Morgan Kaufmann, 1997, pp. 271-280.

W3C, (2010). World Wide Web Consortium website: http://www.w3.org/

Young, S. (2004). ATK: An application Toolkit for HTK, version 1.3", Cambridge University, January 2004

18

Adapting Prosody in a Text-to-Speech System

Janez Stergar[1] and Çağlayan Erdem[2]
[1]*University of Maribor, Faculty of Electrical Engineering and Computer Science*
[2]*Siemens Corporate Technology, Dept. CTIC 5*
[1]*Slovenia*
[2]*Germany*

1. Introduction

The requirements of the evolving information communication technologies (ICT) place new demands on text-to-speech (TTS) systems. The modern high quality TTS system has to be capable of fast and high-quality adaptation to a new language, voice or even expressive speech. Thus adaptation to new voices with different prosodic characteristics is desired.

In this chapter a survey of recent and past approaches of prosodic processing in text-to-speech synthesis will be discussed.

Regardless of the different approaches which have been proposed ranging from generating prosody by rule to huge databases covering almost all prosodic patterns of a specific speaker there is clearly still much work to be done (van Santen et al., 2008).

Automatic learning techniques seem to offer the fastest solution in adapting a TTS system to a new language, voice or a new application. They allow automatic extraction of specific features (e.g. non-uniform unit selection, prosodic regularities extraction) from an appropriate database of natural speech. Such techniques depend on the construction of a large pre-processed corpora (properly segmented, labelled with appropriate prosody labels, etc.). Despite the overall impression that TTS is an inferior task compared to speech recognition, TTS research and development community was not able to produce massive series of consumer products since the early 80es (Dutoit, 2008). Since then a broad spectrum of systems has been developed and successfully implemented – prosody was one of the major tasks to tackle in such systems.

The term "prosody" covers a wide range of features characterizing "the musical qualities" of speech, including phrasing, pitch, loudness, tempo and rhythm. A number of studies suggest that prosody has a great impact on the intelligibility and naturalness of speech perception. Despite the fact that synthesized speech is nowadays mostly intelligible and in some cases sounds undistinguishable from human speech, it still lacks the flexibility and appropriate rendering of expressivity in the synthesized voice.

Text-to-prosody systems based on the use of prosodic databases extracted from natural speech are a key point for development of new TTS systems. One of the major problems in TTS synthesis consists in the automatic generation of natural and intelligible prosody. Therefore the preparation of suitable speech-corpora for automatic prosodic feature extraction is essential.

The pre-processing and labelling in the TTS front end can be performed either automatically or by hand. While automatic labelling can be less accurate than hand labelling, the latter is very time consuming (sometimes also inconsistent). However in some processes, such as segmentation to non-uniform units, which are essential for concatenative TTS synthesizers and verification of automatically labelled data, expert guided procedures can't be avoided. On the other hand many adaptation tasks can be realized automatically or semi-automatically.
In the following sections we will discuss the main three approaches in tackling prosody in a TTS system:

a. the rule based approach,
b. the statistical approach and
c. the minimalistic approach, using as-is prosody in unit selection process minimizing manipulation efforts of units.

We will emphasize the data-driven (corpus-based) approach of extracting prosodic features and focus on the design of a database. We will discuss the basic procedures in the design of the final corpus and steps taken for a suitable corpora preparation and adaptation of learning modules included in the TTS back end. Also all major processes in the TTS front end will be emphasized – the process of labelling the corpus with appropriate tags (e.g. symbolic prosody labels) and the construction of suitable modules for prediction (modelling of different labelling categories). Nevertheless the most important part the adaptation of the TTS back end will also be discussed.
A selective method for classification of different symbolic tags will be introduced and a NN structure based on auto-associative classificators for modelling prosody presented. The symbolic information in this stage is ported to the final and most important module – the module for acoustic modelling. In this section state of the art NN structures suitable for adaptation to a new language without language expertise were implemented. Our discussion will be concluded with the evaluation tests performed on the adapted multilingual TTS system for Slovenian language.

2. Prosody and TTS

The conversion of text to speech can be described as essentially a two-stage process (Fig. 1) – these are an analysis stage, which derives a linguistic structure from the input text and generation stage, where linguistic structure is used for speech synthesis including the generation of intonation, rhythm and so on.

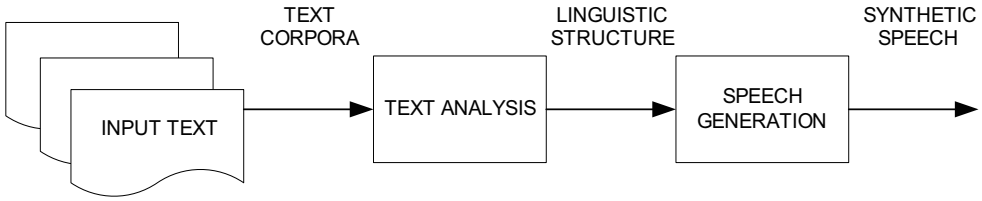


Fig. 1. Text-to-speech process.

The result of text and linguistic analysis in the text front end stage is detailed linguistic information about the structure of the input text, in particular, syntactic, lexical, and phonological with some semantic information. Only the proper choice of prosodic

parameters given by sound duration and intonation contours enables a TTS backend to produce natural-sounding, high quality, synthetic speech (Edgington et al., 1996).
One of the major problems in text-to speech synthesis system consists in the automatic generation of a natural and intelligible prosody. There are two main approaches to the prediction of prosodic structures - rule-based and stochastic.

2.1 Rule based prediction

Many of the rule-based approaches stem from early work on performance structures based on experimental data, such as pausing and parsing values. This work sought to account for the disparity between linguistic phrase-structure theories and actual performance structures produced by humans, and focused on recreating the pause data of several analyzed sentences from syntax (although they claimed that their method could easily account for other prosodic features). The central tenet of the work was that prosodic phrasing is a compromise between the need to respect both the linguistic structure and performance aspects of the sentence.
More recent efforts have extended the work on performance structure prediction to the prediction of prosodic phrasing. In this work, the basic rule-based approach is preserved, but other factors are introduced which are considered important for predictive purposes. For example, some of the researchers believe that syntax plays a lesser role in determining phrasing, and those certain prosodic performance constraints, such as length, override syntactic structure. They allow prosodic boundaries to cross-syntactic boundaries under certain conditions, whereas early work was essentially interclausal. Other modifications include counting phonological words rather than actual words when determining node strengths. A phonological word effectively functions as one spoken item, as the internal word-word boundaries are resistant to pausing. Typical examples are determiner-noun word groups, such as the `the + man'. Further extensions incorporate punctuation into the predictive models, or assign more importance to specific features (Edgington et al., 1996).

2.2 Data-driven or stochastic methods

With the availability of large corpora, annotated with prosodic information such as location and salience of pauses, temporal information on duration, etc., the stochastic-based approach will come more to the fore. Recently methods for automatically predicting prosodic information using decision tree models have been described. Generally, decision trees are derived by associating a probability with each potential boundary site in the text, and relating various features with each boundary site (e.g. utterance and phrase duration, length of utterance - in syllables/words - positions relative to the start or end of the nearest boundary location, etc). The resulting decision tree provides, in effect, an algorithm for predicting prosodic boundaries and their salience (i.e. relative importance) for new input texts.
It is interesting to note that evaluations of both rule-based and data-driven methods recently showed that similar results are achieved (Edgington et al., 1996).

3. The database construction

The first step towards a data driven learning process is the design of speech corpora suitable for (in our case concatenative) TTS synthesis. Also many other aspects in speech synthesis have to be considered e.g. modelling of speech specific attributes (e.g. prosody, emotions,

etc.). One has to consider that literate native speakers have to deal with many important production tasks in the interpretation procedure of read speech (controlled laboratory speech). Pronouncing words correctly is only part of the problem faced by human readers. In order to sound natural and to sound as if they understand what they are reading, one must also appropriately assign prominence to some words, and de-emphasize others. It is unavoidable to 'chunk' the sentence into meaningful (intentional) phrases. Appropriate fundamental frequency (f0) contours have to be chosen and control of certain aspects of voice quality performed. One should also pay attention that a word should be pronounced longer if it appears in some positions in the sentence, that if it appears in others, since segmental durations are affected by various factors, including phrasal position (Sproat & Olive, 1995).

Despite the trend in high quality concatenative TTS systems, to gather as much as possible material (Campbell & Mokhtari, 2003), our goal was to make a compromise between quality and financial means available. We strove for a solution of all in one corpus with the goal to use only one large enough for all major data-driven tasks needed in adaptation of the multilingual TTS system.

3.1 The corpus

The corpus consists of app. 1200 sentences in the Slovenian language (orthography), which equals approximately three hours of speech. The selection of the corpus text was designed to ensure good coverage of the phones in the Slovenian language; therefore clauses were gathered and included from different text styles (e.g., literature and newspaper texts).

The main concern in corpus design was towards optimized suitability for concatenative speech synthesis (the best coverage of elementary segments). No intentional balancing of clause types was performed (declarative – interrogative – exclamations), dialogue context and syntax were not considered, and no semantic analysis was performed since only isolated sentences were included. Prosody was not the first concern for text selection.

The whole corpus was determined using a selection of clauses from a 31 million word corpus in the Slovenian language from e-newspapers, e-literature, the WWW or CD's. The major parts of the clauses covered daily-published news and Slovenian literature; the minority consisted of clauses taken from Slovenian poetry.

In the first step sentences not shorter then 15 and not longer than 25 words were pre-selected from the major corpus. Then, four different text corpora were generated and analyzed statistically (approximately 5000 sentences per corpus). The selection of sentences for the final corpus was based on a two-stage process. In the first stage an analysis based on statistical criteria was performed. In the second stage the final text was chosen based on the results of the first stage. In the following the two stages are described briefly.

After the grapheme-to-phoneme conversion the statistic analysis of corresponding units (mono-phones, diphones, …) generated in a non-uniform unit generator was performed at the sentence level for each of the four corpora in a separate module. The analysis module scans all non-uniform units and determines how frequently each unit appeared. The obtained statistics mirror the non-uniform unit richness and unit structure of all clauses for the corresponding text corpora (Rojc & Kačič, 2000).

After the described statistical analysis of the four different text corpora the final corpus was generated for each of the four. The criterion for the final text filtering was based on monophone, diphone, triphone and fivephone (non-uniform units) richness. Considering

comprehension and frequency of units, a careful elimination of sentences was performed. Clauses with poor unit comprehension and unit duplicates were eliminated. In the final corpus 1200 sentences remained (Rojc & Kačič, 2000).
The statistically analyzed corpora had similar unit statistics, although the distribution of units was not the same. Three of the four corpora included many foreign names (clauses gathered predominantly from newspapers) that we replaced with Slovenian ones, essentially not influencing the statistics of non-uniform units. The corpus with the minimum changes of the non-uniform units after foreign name replacements was chosen as our final corpus.

3.2 Audio recordings

The audio database recordings were created in a studio environment with a male speaker reading aloud isolated sentences in the Slovenian language (fs = 44.1 kHz, 16 bit).
Because the speaker was a professional radio news speaker, the speech contained no disfluencies (i.e., filled pauses, repetitions and deletions) although there was some evidence of hesitations in the form of pauses and lengthening. Compared to the German used in Müller et al., 2000_b, the percentage of hesitations differed significantly (<0.5% German, >15% Slovenian). The stated comparison was estimated after statistical analysis of B9 tags inserted by the labellers in the procedure of phrase breaks labelling.

3.3 Phonetic transcription

The phonetic transcription was managed using a two-step conversion module. The first step is realized with a rule-based algorithm. The second step was designed with a data-driven approach (NN were used). The module was designed for the support of two approaches in grapheme–to–phoneme conversion. The first part was intended for those cases in which no morphological lexica were available. The first rule based stress assignment was applied, followed by a grapheme-to-phoneme conversion procedure. The step of stress marking before grapheme-to-phoneme conversion is very important for the Slovenian language, since it very much depends on the type and place of stress. If the phonetic lexicon is available, a data-driven NN approach, represented by the second part in the module, can be used. In the proposed data-driven approach, a phonetic lexicon was used as data source for training the NN (Rojc & Kačič, 2000).
The data preparation, generation of the training patterns and the training of NNs were done completely automatically. The transcription was performed in two steps. In the first step the graphemes were converted into phonemes, and syllable breaks were inserted in the phoneme string. In the second step the stress marks were inserted. The problem of how to perform mapping between graphemes and phonemes by generating training patterns for NNs (NN) was solved as proposed in Hain, 1999. For both NN tasks we applied a multilayer perceptron (MLP) feed-forward network with one hidden layer. As a learning algorithm, the back-propagation algorithm was chosen.
Pronunciation was derived from the IPA Alphabet. In order to represent the IPA symbols in ASCII characters the SAMPA format is widely used. In our grapheme-to-phoneme conversion module the SAMPA phonetic transcription symbols for the Slovenian language were used (Kačič & Zemljak, 1999).

3.4 Phonetic segmentation

The spoken corpus was phonetically transcribed using HTK. Along with standard nomenclature, two special markers were used for pauses between phonemes. “sil” denotes

the silence before and after a sentence. "sp" denotes the silence between words in a sentence. Both were determined with a one-state HMM and all phonemes with three-state HMM in the HTK environment. The 'sil' and 'sp' tags were used in the semi-automatic process of phrase breaks labelling.

3.5 Part of speech tags

The text corpus was hand-labelled using 13 different classes of part-of-speech tags (POS) among which two were used for punctuation marking (end/intermediate). All tags were combined in an environment where tracking and correcting tags was simplified for the labellers. First an existent environment for reviewing in Microsoft Word was used, but due to stability problems it was replaced with a public domain editor using macros. The first environment was promising with the ability of different reviewing marks with accompanying time stamps and user (different labellers) specific tags (colour-marked).
Compared to the POS tag set in the German corpus used in Müller et al., 2000_b, the POS tag set for the Slovenian corpus is smaller. The difference in size occurs because the Slovenian corpus is hand-tagged, and no reliable tagger currently exists for a large POS tag set and the possibility of reliable automatic POS-tagging.

4. Symbolic prosody labelling

As automatic approaches usually depend on some manual examination and eventual corrections (verification) it seems to be appropriate approaching the problem of labelling with a semi-automatic method.
In our approach of corpora preparation we designed a graphical environment, which we applied for semi-automatic phrase break labelling. The tool was planed to simplify the labeller decisions and support the classification of different classes of breaks.

4.1 Phrase breaks labelling

Since no inventory for symbolic prosody breaks labels is defined for the Slovenian language, it was decided to use labels similar to those used in Kompe, 1997 and Mihelič et al., 2000. Thus the prosody break labels are determined through acoustic perceptual sessions, and the text was labelled speaker dependent (the decisions on labelling were made exclusively on perceptual criteria). The following inventory of prosody break labels was used for labelling the corpus:

- B3 – full intonational boundary with strong intonational marking, often with lengthening or change in speech tempo (we'll refer to that label as a major break);
- B2 – intermediate phrase boundary with weak marking (we'll refer to that label as a minor break);
- B9 – irregular prosodic boundary, usually disfluencies at hesitations, repairs etc. ; and
- B0 – normal word boundary.

The acoustic prosodic boundaries were determined by boundary indication, listening to audio files and visual output (pitch and energy) from specially designed graphical tool.

4.2 Labeling of prominent words

It was decided to distinguish between word accent, phrase accent and sentence (utterance) accent. Word accent is carried by a word emphasized through perceptual prosodic accent or

pitch accent, where phrase accent by our definition is carried by a word most prominent within a phrase comprised of one or more accentuated words. The third accent defined in our inventory is the so-called sentence accent, which is (eventually) carried by a word most prominent in the considered sentence (it is not necessary that a distinction of the so-called sentence accent can be made between words being prominent). The classification of specified accents is a complex matter; therefore, an inventory adequate to distinguish among the three accents was chosen. However, the performed experiments concentrated only on the reduced categories defined in our accent-labelling inventory using two labels (AC = accented, NA = not accented). In the used inventory a phrase is a sequence of words within B2/B3 boundaries.

Distinction between accented and non-accented words was done within a phrase comparing syllable pitch envelope and normalized syllable mean average pitch changes (normalized on syllable mean average pitch changes for the concerned sentence). Energy and mean energy for syllables in each word were also considered. Through acoustic-visual sessions with our graphic tool also a classification in special cases was made where, depending on the accent type, the accented syllables had low average pitch compared to the sentence average. A German-like TOBI intonational description scheme was used for intonational marking (Benzmüller & Grice, 1999). Word prominence is classified according to four classes:

- EA = Emphatic accent,
- PA = Primary accent,
- SA = Secondary accent, and
- NA = No accent.

We consider primary accent to be assigned to normally accented words – words perceptibly most prominent within a phrase (lexical stress). Usually one or more words within a phrase carry a primary accent. We consider the secondary accent to be conveyed by an accentuated word within a phrase, not carrying a primary accent. Finally, the emphatic accent is reserved for accented and (lexical) non-accented words that are perceived as extremely stressed relative to other words or are carrying an emphatic function.

4.3 Selection of phrase breaks

A tool intended to help the labeller (novice or expert) to make decisions about prosody breaks within each sentence was designed (Stergar et al., 2003). The tool indicates possible prosody boundaries, which depend on the segmented pauses in spoken corpora. Prominent words are also indicated.

Experiments on multilingual databases (3 languages) have shown that the strategy of segmenting the speech signal with pauses yields a significant improvement in annotation accuracy (Vereecken et al., 1997). Therefore syllable and word boundaries were marked with vertical lines adding overview clearness, and *B* marks for symbolic prosody boundaries were inserted in the sentence concerned.

The designed tool indicates markers for prosody boundaries taking phonetic segmentation of pauses into account. Yet considering only the duration of silence between phrases, it indicates the position of prosody boundaries. The decision for break indication is made by comparison with a specific threshold. This threshold can be changed manually and tuned according to a specific speaker (Stergar & Hozjan, 2000). However, boundaries indicated by intonational marking or lengthening (without a pause) must still be hand-labelled.

4.4 Selection of prominent words

Two classes of prominence on word level were defined (Stergar & Horvat, 2003):

- perceptual prosodic accents (words being emphasized by stress) and
- pitch accents (words being emphasized by pitch movements).

Our aim was the selective detection of both classes automatically. The hand labelling of prominent words of our database is in progress but is due to a very time consuming process proceeding very slowly.

The first acoustic parameter involved in our experiments was band-pass filtered energy. We used a classical FIR with frequency bounds between 500 – 2000Hz. Experiments in Tamburini, 2002 for Italian and Sluijter & van Heuven, 1996 for American English and Dutch (both for male speakers), showed that this band of high frequencies is the most suitable. For every utterance we computed RMS of the band-pass filtered energy. Energy variations across different utterances were reduced with normalizing every syllable with mean syllable energy over the concerning utterance.

The second acoustic parameter was fundamental frequency – f0. As the extraction of the pitch contour is a delicate task we used a successful scheme for f0 estimation. Therefore we used a robust algorithm for periodicity detection in the autocorrelation domain, suggested in Boersma, 1993.

Additionally we processed every utterance and computed a measure for pitch changes – pitch dynamics (f_D) – for every syllable (Hozjan & Stergar, 2002):

$$f_{D_j} = \sum_{i=1}^{N} \left| x_{i+1} - x_i \right| \tag{1}$$

where j is indexing the current syllable and i the concerned sample.

The prominent syllables (words) were automatically classified according to the proposed statistical threshold selection criteria for perceptual and pitch prosodic accents (Stergar et al., 2003).

5. Acoustic prosody modelling

Data-driven prosody generation modelling by NNs was established first in Traber, 1992 and became state-of-the-art technique. The commonly used NN architecture is the multi-layer-perception with a direct recurrence (Traber, 1992) or without (Haury & Holzapfel, 1998; Erdem et al., 2000). The advantage of modelling with NN is fast and easy adaptation to new languages, speakers and speaking styles. The used TTS system has a language and speaker independent core with external knowledge sources like lexica and NN modules for special tasks. The f0-contour generation module is such an external adaptable source.

5.1 The concept of NN adaptation to a new language

Building up a NN for such a task for the first time or during the adaptation to a new speaker or even to a new language is a delicate task. There is a dilemma how many inputs available to use trying to avoid high input dimensions of a NN. An over-fitting problem of NN's with a high input dimension (which brings lower generalization abilities) is known (Prechelt, 1998). This problem was overcome using expert knowledge e.g. knowledge about the importance of the input parameters or time consuming heuristic approach e.g. testing different input parameter constellations to get the best performance (Sonntag et al., 1997).

We propose a parametric weight decay method, which enables to overcome the difficulties of data-driven techniques with fast adaptation without language expertise. This parametric weight decay systematically analyses the input vector of a NN. In an additional pre-processing unit of the original NN a diagonal matrix to a pre-processing cluster propagates the input parameters. The weight decay concept is applied only to these diagonal elements. These elements represent a weighting of the according input, which then allows an evaluation of input parameters.

The weight decay concept helps training NN models with reduced degree of freedom by adding a penalty term to the error function (Eq. 2). The first term $E_0(w)$ is the original error function and the second $W_P(\lambda, w_i)$ is the standard weight decay penalty function given by:

$$W_P\left(\lambda, w_i\right) = \frac{\lambda}{2} \sum_{w_i} w_i^2 \tag{2}$$

where

λ denotes a penalty scaling factor,

w_i denotes the weights of the NN the penalty term is applied on.

During minimization small values for the weights w_i of the weight vector W were preferred as high valued weights lead to big penalties (Eq. 3). The penalty term used encourages smoother NN weight mappings (Bishop, 1995).

$$E\left(w^j\right) = E_0\left(w^j\right) + \frac{\lambda}{2} \sum_{w_i^j} \left(w_i^j\right)^2 \rightarrow \min_{w_i} \tag{3}$$

where

j denotes the number of the training epochs of the NN penalty scaling factor.

The weight adaptation is performed using the adaptation parameter η, which controls the step size during NN training (Eq. 4).

$$w^{j+1} = w^j - \eta \frac{\partial E\left(w^j\right)}{\partial w^j} \tag{4}$$

Substituting the second term in Eq. 4, with its partial derivative leads to:

$$w^{j+1} = w^j - \eta\lambda w^j - \eta \frac{\partial E_0\left(w^j\right)}{\partial w^j} \tag{5}$$

The weight vector w_{j+1} for the next epoch in Eq. 5 is computed as the difference of the prior weight vector w_j and the partial derivative of the error function (back propagation) as applied in Müller et al., 2000_a and Hain & Zimmermann, 2001. The final weight adaptation is performed with the modified weight vector w_j, where the weight decay term is extended with a parameter p in Eq. 6:

$$E\left(w^j\right) = E_0\left(w^j\right) + \frac{\lambda}{p} \sum_{w_i^j} \left|w_i^j\right|^p \rightarrow \min_{w_i}, \quad 0 < p \leq 1 \tag{6}$$

and with a modified penalty term with its weight adaptation in Eq. 7:

$$w^{j+1} = w^j - \lambda\eta sign\left(w^j\right)\left|w^j\right|^{p-1} - \eta\frac{\partial E_0\left(w^j\right)}{\partial w^j} \tag{7}$$

In the following this modified weight decay function will be referred as *p-WD* (according to the term of the introduced parameter p). There are different ways to implement the weight decay within NN training, as one might apply the penalty term to all weights within the NN or to special connection areas. The input vector x is propagated by the *diagonal matrix* $W_{diag} \in R^{l\times l}$ to the pre-processing layer, which has a *tanh* activation function (Fig. 2). This diagonal matrix is the only connection that utilizes the penalty term in Eq. (6). Due to this fact the order of vector element x_i' is defined by the product of the *diagonal matrix* W_{diag} and the input vector x_i. Therefore w^i_{diag} gives a weighing of the input x_i. w_i is bound to the interval [0,1]. After pre-processing of input data the weighed inputs are propagated to the original NN with n hidden neurons and m outputs. The original NN will be explained in detail in the after going paragraphs. It is important to initialize W_{diag} with equal values, which are incremented or decremented according to the weight adaptation in Eq. (7) during training. This type of realization of weight decay aims at:

- Outlier cancellation: the asymptotic behaviour of *tanh* is used in the pre-processing layer as a delimiter. An element of W_{diag} growing too high is limited to the interval [-1;1] at the output side of the pre-processing cluster. With this precaution measure inputs are avoided to dominate the mapping. All inputs are limited to the interval [-1;1].
- Soft input pruning: elements of W_{diag} being pushed to zero are considered to have no significant influence on the NN model. This means that the corresponding input contains no important information for the training task with the used database. Inputs being close to zero might be removed afterwards by introducing a threshold and omitting inputs with values in W_{diag} below that threshold. We obtain soft input pruning, as there is no ultimate decision made during training. Unimportant inputs are faded out.
- Separation into subtasks: The original task and the input feature analysis are solved in a parallel manner. The feature analysis does not need a further training phase.

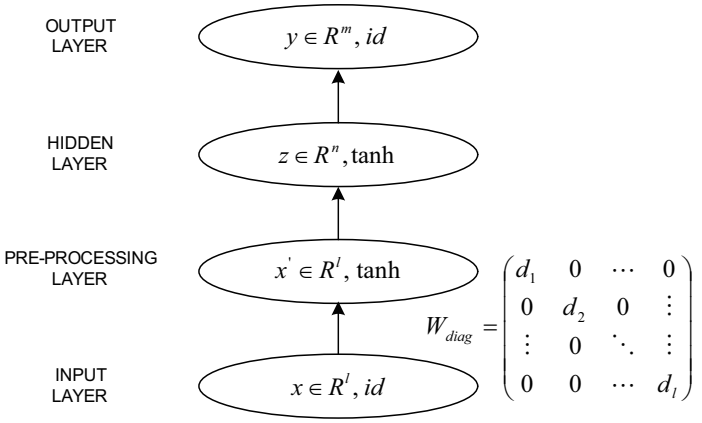


Fig. 2. Realization of p-WD.

Dealing with highly correlated input features the application of *p-WD* should be preferred, as it helps to select one of the highly correlated features in contrast to standard weight decay. Weight decay does not select one of the highly correlated inputs. This different selection property of *p-WD* is exemplified using a NN with two highly correlated inputs. This can be formulated in Eq. (8).

$$x' = \tanh\left(w_1 x_1 + w_2 x_2\right) \tag{8}$$

If we presume that inputs x_1 and x_2 are highly correlated we can rewrite Eq. (8) into:

$$x' = \tanh\left(\left(w_1 + w_2\right) x_1 + 0 \cdot x_2\right) \tag{9}$$

The penalty terms for the standard weight decay are given by Eq. (10). The left side of the inequation according to Eq. (8) and the right side according to Eq. (9):

$$\frac{\lambda}{2}\sum_w w^2 \quad : \quad w_1^2 + w_2^2 \geq \left(w_1 + w_2\right)^2 + 0^2 \tag{10}$$

Using the Lagrange multiplier method and the penalty terms for Eq. (8) and Eq. (9) we can state the final inequation for the modified p-WD with the introduced parameter p as follows:

$$\frac{\lambda}{p}\sum_w |w|^p \quad : \quad \left|w_1\right|^p + \left|w_2\right|^p \geq \left|w_1 + w_2\right|^p + 0^p \tag{11}$$

Through the p-WD penalty function we achieve a minimized input feature set as the right side of Ineq. (11) is smaller than the left side. In experiments described in Erdem & Zimmermann, 2002_c, the stated behaviour of p-WD has been observed and the optimal parameter p determined.

5.2 f0 generation

The p-WD method presented in the foregoing section is implemented in the f0-contour generation module. The utilized NN has to map input parameters to an appropriate f0-contour. Regarding the syllable the mapping is performed to four f0-contour parameters (Fig 3). The solid line depicts a f0-contour on the syllable level. These contours are parameterized (dashed line) by four values: f0-maximum (p1= f0_max), f0-maximum position (p2= f0_maxPos), f0 at syllable start (p3= f0_Start), and f0 at syllable end (p4= f0_Stop). For the contour parameterization a maximum based description is used (Heuft et al., 1995), which mainly defines that f0-contours on syllable level for non-tonal languages can be described by a rising on the first part and a falling on the second part of the syllable.
The mentioned parameters p1, p2, p3, and p4 are the outputs y = {p1, p2, p3, p4} of the NN respectively (Fig. 4). Hence the dimension of the output cluster m = 4.
f0-contours are known to be influenced by long-term features (the sentence type), breath and local stress intention. The input parameters must contain information concerning local and global characteristics (symbolic prosody tags). For a good mapping it is also important to provide contextual information of the syllable. Due to computation reasons the context window length was chosen to be seven to the left (past) and seven to the right (future) of the syllable with the exception of the linguistic categories. The following input features are presented to the NN to solve this problem on the syllable level for each context unit:

- Phonetic information: The phonetic structure of a syllable to be processed is coded here. The vowel is presented as a one out-of-n coded input using the Slovenian SAMPA phoneme set. Neighbouring phonemes of the vowel are given in four classes (plosive, fricatives, nasal and liquids) and also as a one-out-of-n coded input in a symmetric context window of four phonemes.
- Positional information: Continuous positional information gives time distances of the syllable and its vowel. Discrete information denotes whether this syllable is an initial, medial or final one within the sentence, the phrase, and the word.
- Stress information: Flags denote the stress type of a syllable within the word and the phrase.
- Linguistic categories: The used linguistic category set consists of 14 tags, which are one-out-of-n coded and presented in a context of 3 to both sides.

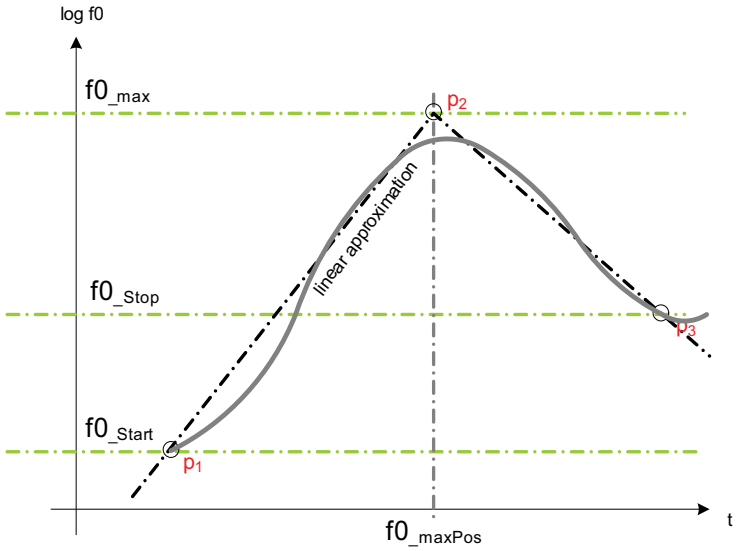


Fig. 3. Maximum based parameterization of f0-contours.

Hence by this input constellation the input dimension of x according to Fig. 2 and Fig. 4 is in the range between 500 and 600 Hz. The p-WD technique was applied to recordings of three hours of a Slovenian news speaker reading gathered isolated sentences from a large corpus as described in the foregoing section.
The patterns for training (80%) and testing (20%) were separated. A validation set of (20%) was selected randomly from the training set. The introduced parameter p in the weight decay penalty term of *p-WD* was optimized by experiments with varying parameters p (p=[0.1 - 2.0; step 0.1]). The optimum parameter was found to be p = 0.6 (Erdem & Zimmermann, 2002_a). This tuned NN module was then used to analyze the inputs and optimize the input feature selection.

6. Symbolic prosody modeling

Which parameters are the most relevant for symbolic prosody label prediction remains an open research question. A carefully chosen feature set can help to improve prediction

accuracy; however, finding such a feature set is work-intensive. In addition, linguistic expert knowledge can be necessary and the feature set found can be language and task dependent. A feature set that is commonly used and seems to be relatively independent of language and task is part-of-speech (POS) sequences (Müller et al., 2000_b).

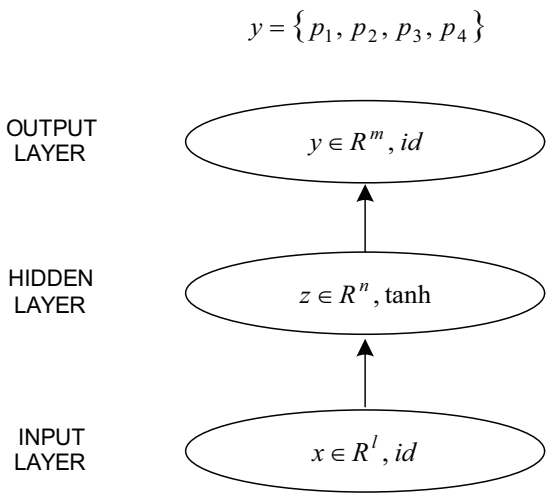


Fig. 4. NN structure for f0-contour modelling.

POS sequences of length four to the left and right of the position in question were used. For input to our prediction model the POS sequences were coded with a ternary logic (–1 for a non-active node, +1 for an active node, 0 for not valid).
Thus, for each POS tag a vector was obtained with a dimension of the size of the tag set. The size of our tag set was 13. Using a POS sequence length of four to the left and right for the Slovenian language, we achieved m = (4+1+4) * 13 = 117 dimensions.
The dimension of the applied input vector as well as tag set is similar to the German language prediction tests as reported in Müller et al., 2000_b where a tag set of length 14 was used.

6.1 Autoassociative NN classifier

We used a new approach of symbolic prosody tags prediction from POS with a NN structure based on autoassociators introduced in Müller et al., 2000_b. With the used architecture we minimized the problem of unbalanced information flow between the forward and backward path where many inputs are compressed into a single number for classification error. The architecture consists of two stages; STAGE 1 and STAGE 2 (Fig. 5).
The first stage consists of k different autoassociator models for k different classes (e.g. k=4 for B1, B2, B3, B9). Each model is trained only with data from the class it represents. The m-dimensional input vector x is mapped onto n-dimensional vector z, with n<<m. The NN are trained with the goal that the output vector x′ recovers as accurate as possible the original input x.
Thus an intermediate representation z of the data in a lower dimensional space is achieved with the compression of x via the matrix w1 and hence decompression of z via matrix w2. After training for each autoassociator a reconstruction error eREC is computed. The distance

The distance measure eREC = (x-x′)² is achieved through a squaring activation function of the upper cluster (Fig. 6, right) considering the difference between input x and x′ achieved using a negative identity matrix -id. The result is high dimensional error information as input into the classifier.

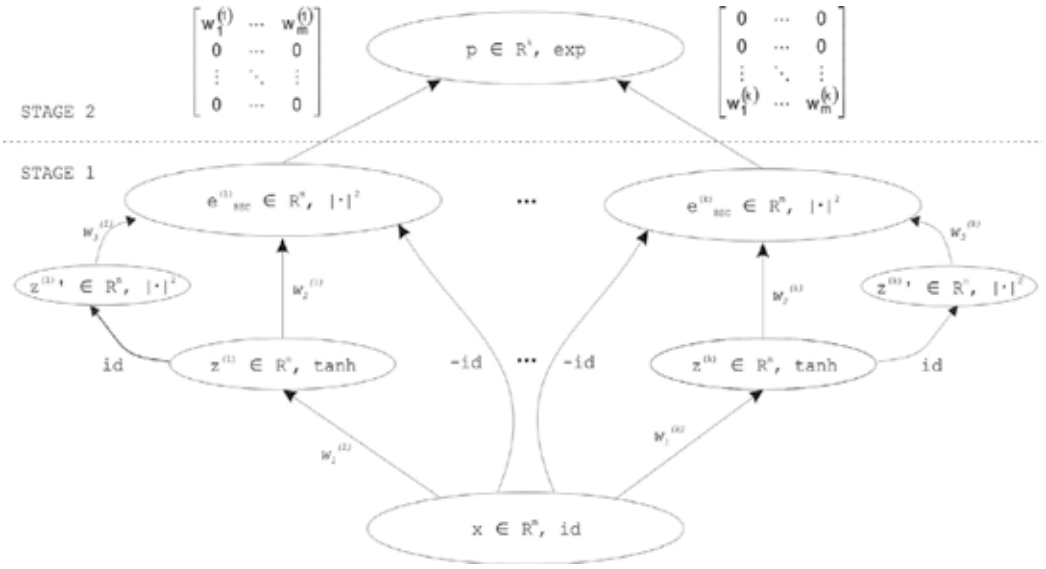


Fig. 5. Autoassociative NN classifier.

The performance of the autoassociator can be improved with augmentation of the coordinate transformation $x' = w_h \cdot tanh(w_i \cdot x)$ additionally taking the squared representation of z into account (Fig. 6, right).

$$p_i = \frac{e^{\left(x - w_2^{(i)} \tanh\left(w_1^{(i)} \cdot x\right)\right)^T diag\left(w_1^{(i)}, w_2^{(i)}, \ldots, w_m^{(i)}\right)\left(x - w_2^{(i)} \tanh\left(w_1^{(i)} \cdot x\right)\right)}}{\sum_{j=1}^{k} e^{\left(x - w_2^{(i)} \tanh\left(w_1^{(i)} \cdot x\right)\right)^T diag\left(w_1^{(i)}, w_2^{(i)}, \ldots, w_m^{(i)}\right)\left(x - w_2^{(i)} \tanh\left(w_1^{(i)} \cdot x\right)\right)}} \quad (12)$$

In STAGE 2 these detailed error information is used to determine which class (model) a given pattern on input x probably belongs to. The classifier is a NN that calculates the class conditional probabilities $p_i = p(x \in class_i)$ from the reconstruction error vectors of the different autoassociator models in Eq. (12). The experimental results confirm our assumption that the characteristics of the different classes can be captured by such autoassociators (Müller et al., 2002).

Our tests of phrase break prediction were performed with limited labelling material available (Stergar et al., 2003), app. ½ compared to Müller et al., 2000_b. The results are comparable to those for German and English (Müller et al., 2000_b; Black & Taylor, 1997). For the prediction of breaks (B correct), the results are equivalent to the achieved accuracy prediction of B correct (77.67 %) for German and nearly equivalent to the achieved accuracy

prediction of B correct (79.27 %) for English despite the reduced inventory of clauses used for training.

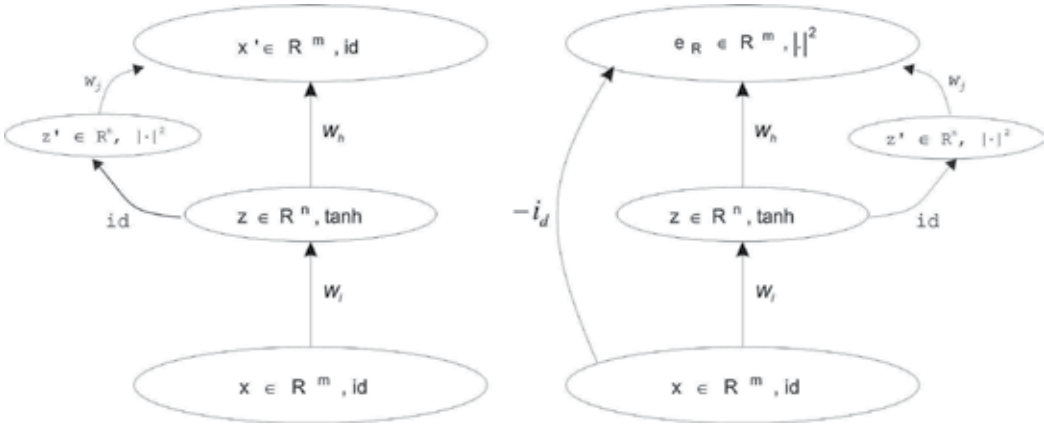


Fig. 6. Left: an autoassociator NN trained for a single class. Right: an autoassociator used for computation of reconstruction error.

7. Duration control

Different methods for duration control have been proposed. In Klatt, 1987, a rule based duration control method was presented, which depending on rules, modifies the duration of a segment by a multiplicative or additive scaling factor. In van Santen & Olive, 1990 the authors differentiate between different duration models, a data-driven method with good generalization abilities is presented in Campbell, 1992 using a NN for the syllable duration control.

Our acoustic prosody module consists of a duration control and f0-contour unit (Fig. 7). In the first approach the introduced duration control module used a classification and regression tree (CART) like method. The parameters for the nodes were derived from triphone clusters obtained by a tree-based clustering algorithm provided by standard clustering within HTK (Holzapfel, 1999). This approach will not be explained in detail. Additionally this statistic method was enhanced by a more sophisticated statistical approach considering larger contextual information on syllable and word level. Nevertheless there is a dilemma between robustness and significance of the statistics as the multilingual aimed TTS system utilizes a restricted speech database (app. 1000-1200 sentences). Therefore a NN approach is employed to solve this dilemma due to its generalization property.

As depicted in Fig. 7 both units' duration control and f0 modelling are modelled by NN. The segmental duration control is handled first – those segmental durations generated by the duration control unit are afterwards used as inputs to the f0-generation unit.

Continuous positional information of syllables is derived from these durations, which are important for the f0-generation task. A segmental duration module has to control the rhythm of a synthetic voice and the known effect of final lengthening.

Similar to the f0-contour prediction task the duration control unit uses left (past) and right (future) contextual information to establish the prediction. The input features of both modules are very similar. They are organized on syllable level for the f0-contour prediction and on phoneme (triphone) level for the duration control task. The state-of-the-art causal

retro-causal modelling was presented in (Zimmermann et al., 2000) for the f0-prediction task. The shortcomings of that architecture are a fix-point recurrence, which causes stability problems during training, and a non-observance of the mentioned structural switching. The causal retro-causal error-correction (CRCEC) NN architecture is used as a basis for the modelling of the duration control task. Different architectures will be presented to overcome these two problems. First basics (finite unfolding, error correction) and new partial retro-causal expansion for the integration of the structural switching are discussed. Also the asymmetric architectures are mentioned which overcome the structural switching of the information flow. We will conclude with the implementation of the asymmetric modelling in the NN applied.

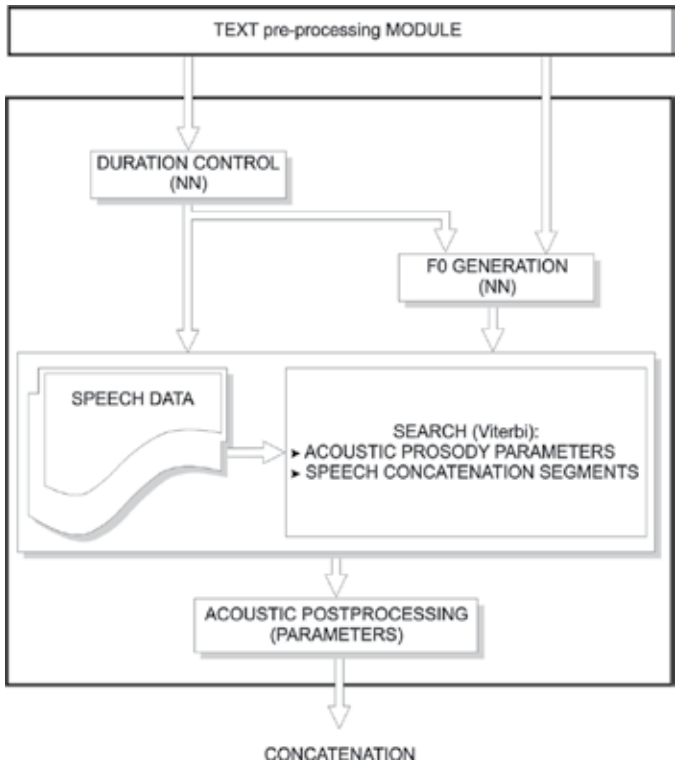


Fig. 7. The acoustic part architecture of the used TTS system.

7.1 The causal retro-causal error-correction architecture

The causal retro-causal error-correction NN architecture (CRCECNN) for one time step (solid lines) is depicted in Fig. 8. This architecture uses shared weights and has a symmetrical extension to the neighbouring time steps (dotted lines). As can be seen there are two different information flows. In the upper part, there is a causal information flow denoted by the *matrix A* carrying state information (s_{t+i}) of the dynamics between neighbouring state clusters. This path allows the mapping of long-term forecasts. The *matrix F* in the lower part gives a retro-causal information flow (r_{t+i}). Within each time step *i* there are two error-correction parts incorporated. Both are coupled by the usage of one output cluster z_{t+i}. The error-correction will be explained using the causal information flow path.

While *matrix B* introduces external information u_t to the system, the *matrix C* transforms the state s_t to its expectation y_t. *Matrix D* propagates the model error (the expectation y_t being compensated by the observation y_t^d) to cluster s_{t+1}. The path:

$$s_t \to C \to z_t \to D \to s_{t+1} \tag{13}$$

allows to map local structures as shocks or short term effects (Eq. 13). The z-clusters represent the output clusters of the NN architecture. In cluster z_t the difference $z_t = C \cdot s_t - y_t^d$ (forecast error) between the expectation of the NN and the observation y_t^d is computed. Note that y_t^d is propagated by identity *Id* to z_t.

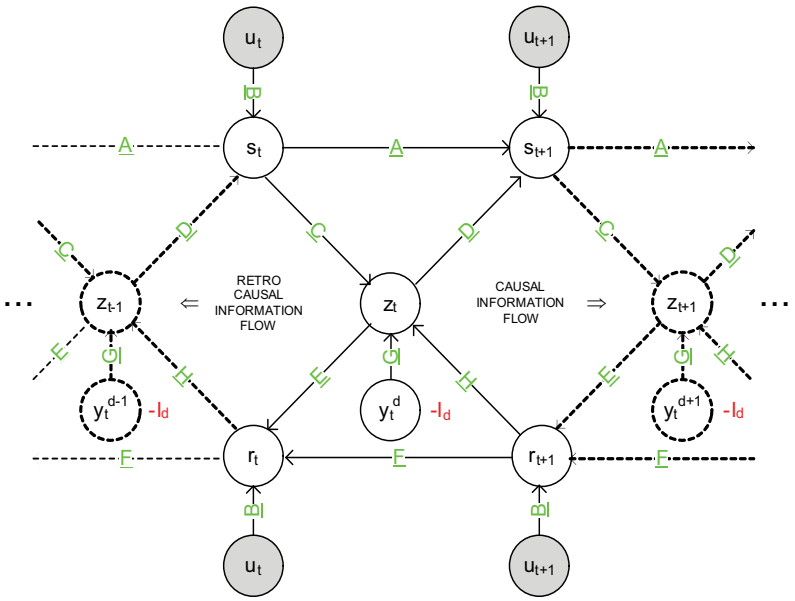


Fig. 8. Causal Retro-Causal structure of modelling.

This difference has its optimum in zero, since this denotes no forecast error having no forecast error results in a perfect description of the dynamics. Therefore the target vector z_t is set to zero during training.

If there is no mismatch between expectation and observation, then no further information is propagated by *matrix D* to state s_t and we almost obtain a simple finite unfolding NN. Existing mismatch delivers further input information to state s_{t+1}. This information is used during training for the adaptation of parameters. By this error correction principle we obtain in z_t an internal vector driving the transition of the system state together with external input u_t and previous states. These internal vectors generate the error flow, computed at each output cluster time step of the unfolding. If the internal autoregressive part coded in *matrix A* and all external driving forces of a dynamics are known, it would be possible to give a perfect description of the dynamical system. But if it is not possible to identify the dynamics due to missing or unknown externals or noise, the last model error is an indicator of the model misspecification. Since the model error is used as a measure of unexpected shocks, the learning of false dependencies is lowered and models generalization ability is improved (Rumelhart et al., 1986; Zimmermann et al., 2002).

Incorporating information flow from the right to the left captures retro-causal dependencies. If this is handled symmetrically over all time steps, this modelling result in fix point recurrences as depicted in Zimmermann et al., 2000. One closed loop is given by:

$$s_t \to C \to z_t \to E \to r_t \to H \to z_{t-1} \to D \to s_{t+1} \tag{14}$$

These fix point recurrences complicate computation of the resulting CRCECNN. Therefore a partial symmetric expansion in the following subsection which results in a *partial CRCECNN* (P-CRCECNN) is an approach to solve the stated problem (Erdem et al., 2002_b).

7.2 The partial causal retro-causal error-correction architecture

The NN depicted in Fig. 9 utilizes shared weights and finite unfolding. The coupling of both information flows is realized by only one output cluster z_t instead of the coupling at each time step within CRCECNN. By coupling that information flows within the present time step this new architecture does not contain fix-point recurrent loops, which might cause instabilities during training. In the following this architecture will be used for further adaptations applying structural switching.

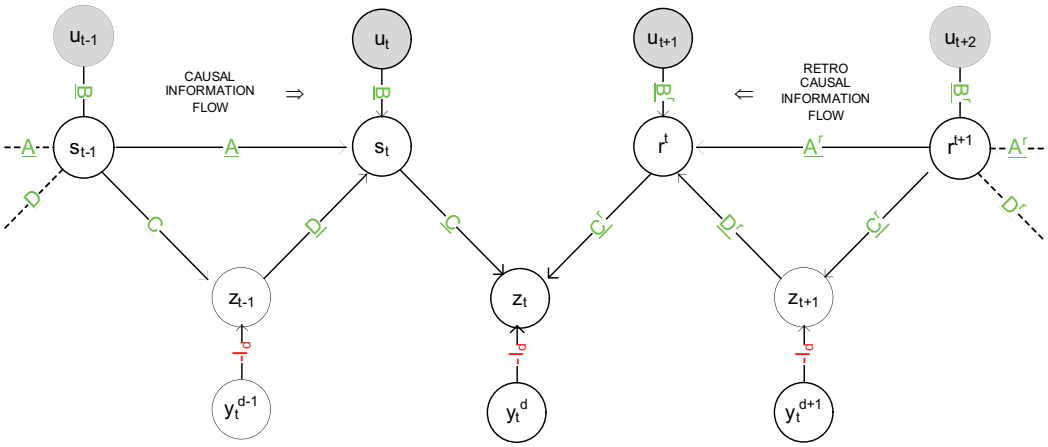


Fig. 9. The structure of P-CRCECNN.

During training all segmental durations modelled as observations are known. But within the application there are no observations available for i ≥ 0, because they are not predicted yet. For i < 0 predictions of the NN are re-utilized as observations. Because of this mismatch between training and application the retro-causal information flow has to be treated in a specific way. In the following to different ways of asymmetric P-CRCECNN are explained which overcome this mismatch. In Fig. 10 the idea of removing connections after training is depicted. The dotted connections *matrix* C^r and *matrix* D^r are trained. So the architecture is the same as depicted in Fig. 9 during training. But within the application connections $\boldsymbol{C^r}$ and D^r are removed. The resulting architecture is then a finite unfolding in time without the error correction principle for the retro-causal information flow during application. The next architecture is established by using finite unfolding in time for the retro-causal path during training and application (Fig. 10, the removed nodes and connections are shown as transparent).

7.3 The implementation in duration control NN module

In the following the application of asymmetric P-CRCRECNN's within the segmental duration control unit of our acoustic prosody NN module is presented. These data-driven methods are applied to recordings of approx. three hours of a Slovenian database as already described (The corpus). The patterns for training (80%) and testing (20%) are separated. A validation set of (20%) is selected randomly from the training set. The used database is the same as applied within the f0-generation task described in the foregoing section (f0 generation). The f0-generation task utilized patterns organized on syllable level – within this task, patterns are organized on triphone level.

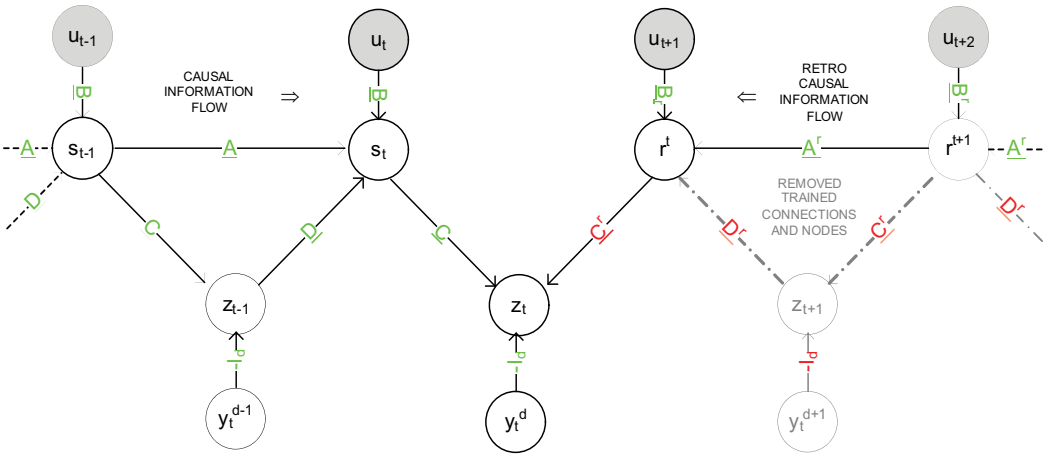


Fig. 10. Topology of modified P-CRCECNN removing trained connections with finite unfolding.

The following information (extracted from the database) is presented to the NN input in a context of seven phonemes to the left and right:

- Phonetic information: with one-out-of-n coding the phoneme index is presented here. A phoneme-set of 45 phonemes is used. Additionally the four phoneme classes (vowel, fricative, nasal, liquid, and plosive) are presented here.
- Positional information: discrete information denotes whether the according syllable is an initial, medial or final one within the phrase and the word. Continuous information is given by the relative syllable position within a sentence and phrase.
- Stress information: flags denoting the stress type of the according syllable are coded here. Four flags present Word level stress (prominence as described in section symbolic prosody). Sentence level stress consists of two stress marks.
- Linguistic categories: a one-out-of-n (set of 13 categories) coded linguistic category (POS) denotes the category type of the according word.

All listed input categories are presented at each time step of the unfolding clusters denoted by u_{t+i}. The according output vectors are modelled as observations and are presented at each time step in the clusters denoted by u^d_{t+i}. Target values for the NN are normalized to ensure an optimized signal-flow during training of the NN due to *tanh* activation function within the causal and retro-causal state clusters. A first normalization of segmental duration is obtained by the mean and standard deviation value from the used triphone classes. A second normalization was necessary to ensure an optimized signal flow during training of

the NN. The mean and standard deviation were derived from the first normalized segmental durations.
For evaluation the trained NN were used to predict segmental durations of sentences that were in the test set. Three audio-files are generated with those predictions that are then used for evaluation.
The experiments showed that the best results are obtained with the P-CRCECNN modified with removing connections after training (Fig. 10). The obtained test showed that 85,6% of phrase breaks were realized with a clear final lengthening (Erdem et al., 2002_b).
As perception is a highly complex process not necessarily modelled appropriately by isolated physically distances, informal listening test were performed. Files generated by re-synthesis utilizing the different presented NN-architectures and the original one were presented to non-expert listeners. They had to judge which of the presented files were most/least pleasant. The target group also had to give a ranking. This ranking was then scaled on a value set from 1 to 5, with 5 denoting the acoustically most pleasant sentence and 1 reserved for unacceptable ones. The asymmetric P-CRCECNN with connections removed after training was evaluated to be most pleasant (average rating of 3,27). This architecture uses the error correction principle within the retro-causal path for modelling local prosodic structures. Hence it seems to help the long term forecast path improving its generalization ability, as this NN performs better than the PCRCECNN with final unfold, which does not utilize error correction. However the long term forecast path within P-CRCECNN structures with final unfold also has the ability to capture short time events.

7.4 Unit selection and Multi-level Viterbi-search algorithm

The unit selection module within the introduced multilingual TTS system uses a robust unit selection method based on syllable prosody parameters optimization (Erdem et al., 2002_c).
First isolated NN predictions of f0-contours (Erdem et al., 2002_a) and segmental durations (Erdem et al., 2002_b) are performed and then these parameters are re-utilized for a search in speech data (corpus) for best fitting of speech segments and acoustic prosody parameters. This search is realized by using a modified multi-level Viterbi-search algorithm that operates on syllable level but explicitly allows higher and lower levels of speech segments in the path search procedure. The selected units (words, syllables and triphones) are chosen from different utterances. Dealing with a limited database it is likely not all specified targets are fulfilled by the units found in the database. Thus a post-processing on the prosody parameters at the selected unit boundaries is necessary.
Regarding the phonetic and perceptive criteria it is crucial to find optimal speech segments. The optimization of physical distances does not necessarily result in a naturally sounding synthesized voice. Perception is a highly complex process, which usually can't be modeled appropriately only by tackling single physical distances between segments. Evaluating the target distance for a candidate unit, or distance between a pair of units to be concatenated, returns only a physical measure of distance, and is not necessarily a reliable indicator of the perceived distortion that may occur (Holzapfel & Campbell, 1998). Therefore nonlinear weighting of the partial suitabilities by multiplying them in order to obtain a global suitability function is applied:

$$S_{\text{global}} = \prod_{n} CSC \cdot \prod_{i} LSC \tag{15}$$

where
CSC = Continuity Suitability Cost (based on phonetic context, prosodic context and acoustic concatenation cost),
LSC = Local Suitability Cost (based on phonetic context, duration, log power and mean f0).
The computation of LSC is based on syllable level prosodic targets, as a syllable level maximum based description (Heuft et al., 1995) of f0-contours for non-tonal languages was used to train the f0-NN. The following targets are utilized for LSC computation (syllable level):

- f0-contour target parameters: p1 (initial f0-value), p2 (maximum f0-value), p3 (final f0-value), p4 (maximum position).
- segmental durations of the triphones,
- power of triphones.

Each LSC has to be in an acceptable range for an acceptable unit candidate. In contrast to a linear weighting by adding each partial suitability (Hunt, 1996) one single partial high mismatch already leads to a low overall S_{global}.
To calculate the different suitabilities a fuzzy logic motivated nonlinear suitability function is used that is composed of two half Gaussians and one constant region. The constant region is dissected to two non-equal regions representing a threshold. Within that threshold, distances to target values have no perceptual influence. Two parameters S_L and S_R control the shape of the Gaussian function.

$$S_{S_X}(t) = e^{\left(\frac{\left(t - t_T + Y\right)^2}{2 \cdot S_X^2} \right)} \tag{16}$$

where
S = nonlinear suitability function,
S_X = Left/Right distance in the Gaussian part of the local suitability cost function LSC; X= {L, R},
Y = threshold region.
S_L and S_R regulate the influence of a special target parameter. A small value of both parameters indicates a sharp criterion for selection, as for the same distance to the target a lower suitability is returned (Eq. 16). All the above mentioned targets (p1-p4) are calculated in this way.
For the CSC calculation a simple but efficient solution was experimentally selected instead of the original time consuming signal processing based on spectral analysis. Experiments showed only minor degradation in speech quality (Erdem et al., 2002_c).

7.5 Post-processing

Dealing with limited speech data (segments) for synthesis makes signal processing on speech elements at concatenation points unavoidable. Therefore we used simple but efficient post-processing on the selected segments prosody parameters. This new method was already applied and tested within the TTS system PAPAGENO for a German male news speaker (Erdem et al., 2002_c). It could be shown that it improves the quality of the used prosody generation module and of the selection process.

It was observed that the used NNs are giving good prosody modelling results within macro prosody. Therefore the general idea of this post-processing is a realignment of the obtained f0-contours according to the run of the f0-maxima of the triangles (Fig. 11). A shift is applied that operates on three levels (word-, syllable, and triphone-level). Values Δ1 and Δ2 give the difference to the according f0-maximum of the syllable or word predicted by the NN. After this shift a jump discontinuity may occur therefore a smoothing is applied using a declining linear function on each side of the jump discontinuities (Erdem et al., 2002_b).

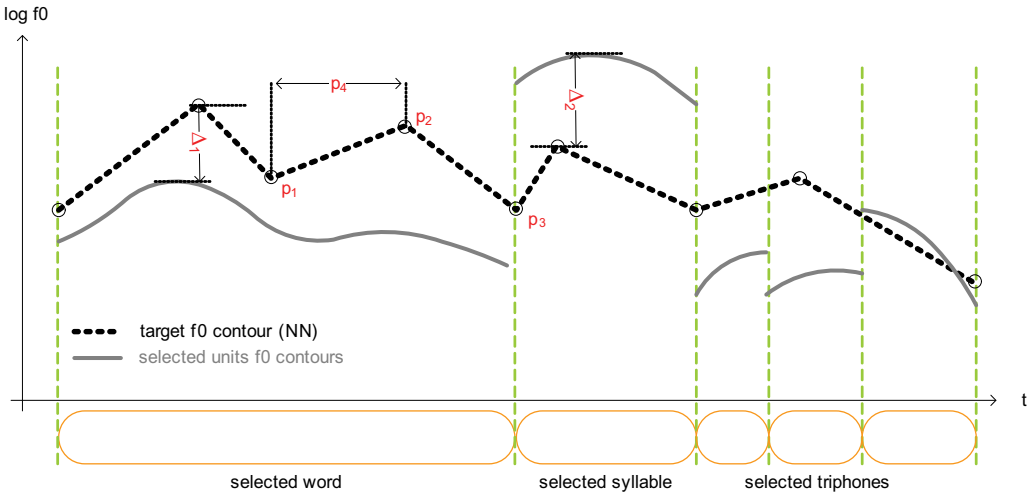


Fig. 11. Contours in selected units with jump discontinuities at unit boundaries.

After the post-processing, the f0-contours are then realized by modifying the speech-elements using a PSOLA like algorithm for speech synthesis.

8. Conclusion

In the foregoing sections we introduced the adaptation of a multilingual TTS system to Slovenian language. In the starting sections we explained our approach in the design of a suitable database used for adaptation of all modules in the used TTS system. We presented the procedure applied in creation of the final corpus, and steps taken for necessary basic pre-processing procedures (grapheme to phoneme conversion, insertion of syllable breaks, syllable stress marks placement, transcription and phonetic segmentation). All steps were performed completely automatically. We introduced a semi-automatic approach in symbolic tags marking for hierarchical prosody modelling used in the acoustical part of TTS system. The presented procedure for phrase breaks labelling is based on HTK tags for silence and is performed semi-automatically. The automatically selected tags were manually verified also other important marks had to be inserted after the process of verification manually. With the introduced approach we accelerated hand labelling and contributed to consistency in the labelling procedure. In comparison to other results our approach shows almost the same or slightly improved prediction performance with 50% less data used for training. However we have to mention the differences (selective labelling) in the labelling procedure used in our approach. In more detail we explained the used adaptable acoustical architecture combined

of four modules. The first module introduced was the duration control NN module. We emphasized its basic structure with the new p-WD method applied. The p-WD method helps to select one of the highly correlated features in contrast to standard weight-decay. Hence through its penalty function we achieved a minimized input feature set. The NN duration control module introduced uses the modified causal retro-causal error correction architecture (CRCECNN). With the introduced architecture the module error is used as a measure of unexpected shocks, the learning of false dependencies is lowered and module generalization ability is improved. The fix point recurrences computation difficulties were solved with the proposed partial CRCECNN architecture. The performed experiments confirmed the suitability of the P-CRCECNN architecture. The problem of finding optimal speech segments was also mentioned. We used an approach of segment selection using a global parameterized non-linear suitability function in combination with a modified multi-level Viterbi search algorithm. Nevertheless due to the fact of a limited database a post processing approach had to be implemented.

The acoustical results of our adapted multilingual TTS system were presented to a group of 20 non-expert listeners. We generated an inventory of 216 test sentences not used for the training or validation process. The test performed during a 3 hour session (approx.) showed that our approach of adapting a multilingual TTS acoustic architecture based on NN architectures is suitable and promising. The average rating (1-5) was good-very good (3,28). We also conclude that the implementation of symbolic prosody tags into the architecture of acoustic modelling essentially contributed to naturalness of the synthesized speech without influencing the intelligibility of synthesized sentences.

9. References

Bishop C. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, ISBN 978-0-19-853864-6, Oxford

Black A. W., Taylor P. (1997). Assigning Phrase Breaks from Part-of-speech Sequences, *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 995-998, Rhodes, Greece

Boersma P., (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *Proceedings 17*, pp. 97-110, Institute of Phonetic Sciences, University of Amsterdam

Campbell N., (1992). Syllable-based segmental duration, In: *Talking Machines: Theories, Models and Designs*, Bailly G., Benoit C. and Sawallis T. R., (Ed.), pp. 211-224, Elsevier Science Ltd., ISBN: 978-0444891150, North-Holland

Campbell N., Mokhtari P., (2003). Voice Quality: The 4th Prosodic Dimension, *Proceedings of the 15th International Congress of Phonetic Sciences,* pp. 2417-2420, Barcelona, Spain

Dutoit, T. (2008). Corpus-Based Speech Synthesis, In: *Handbook of Speech Processing*, Benesty J., Sondhi M., Huang Y., (Ed.), Springer Handbook of Speech Processing, pp. 437-455, ISBN: 978-3-540-49125-5, Springer Berlin Heidelberg

Edgington M., Lowry A., Jackson P., Breen A. P., Minnis S. (1996), Overview of current text-to-speech techniques II – Prosody and speech generation, *BT technology journal*, Vol. 14, No. 1, pp. 84-99, Springer Dordrecht, ISSN 1358-3948, Holland.

Erdem C., Beck F., Hirschfeld D., Hoege H., Hoffman R. (2002_c). Robust unit selection based on syllable prosody parameters, *Proceedings of 2002 IEEE Workshop on Speech*

Synthesis, pp. 159 – 162, ISBN: 0-7803-7395-2, Santa Monica, California, USA, Sept. 11-13, 2002

Erdem C., Holzapfel M., Hoffmann R. (2000). Natural F0-contours with a new Neural-Network-hybrid approach, *Proceedings, 6th International Conference on Spoken Language Processing* (*ICSLP 2000)*, pp. 227-230, Beijing, China, Oct. 16-20, 2000

Erdem C., Zimmermann H. G. (2002_a). A data-driven method for input feature selection within neural prosody generation, *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing ICASSP 2002*, pp. 477-480, Orlando, Florida, May 13-17, 2002

Erdem C., Zimmermann H. G. (2002_b). Segmental duration control by time delay neural networks with asymmetric causal and retro-causal information flows, *Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN 2002)*, pp. 269-274, ISBN 2-930307-02-1, Bruges, Belgium, Apr. 24-26, 2002

Hain H. U., Zimmermann H. G. (2001). A Multilingual System for the Determination of Phonetic Word Stress Using Soft Feature Selection by Neural Networks, *4th ISCA Tutorial and Research Workshop (ITRW) on speech synthesis (SSW4)*, Blair Atholl Palace Hotel, Perthshire, Scotland, Aug. 29 – Sept. 1, 2001

Hain H. U. (1999). Automation of the training procedure for neural networks performing multilingual grapheme to phoneme conversion, *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, vol. 5, pp. 2087-2090, Budapest, Hungary, Sept. 5-9, 1999

Haury R., Holzapfel M. (1998). Optimisation of a Neural Network or Pitch Contour Generation, *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, Vol. 1, pp.297-300, Washington State Convention Center, Seattle, Washington, USA, May 12-15, 1998

Heuft B., Portele T., Höfer F., Krämer J., Meyer H., Rauth M., Sonntag G. (1995). Parametric Description of F0-Contours in a Prosodic Database, *Proceedings of the 13th International Congress of Phonetic Sciences (ICPHS 95)*, Vol. 2, pp. 378-381, Stockholm, Sweden, Avg. 13-19, 1995

Holzapfel M. (1999). HMM based database segmentation and unit selection for concatenative Speech Synthesis, The Journal of the Acoustical Society of America, Volume 105, Issue 2, p.1031, Feb., 1999

Holzapfel M., Campbell N., (1998). A Nonlinear Unit Selection Strategy for Concatenative Speech Synthesis Based on Syllable Level Features, *Proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP 1998)*, Vol. 6, pp. 2755-2758, Sydney Convention and Exhibition Centre, Darling Harbour, Sydney, Australia, Nov. 30 – Dec. 4, 1998

Hozjan V., Stergar J. (2002). Determination of prominence accent of prosodic segments in emotional speech, *Advances in Speech technology: 8th International Workshop*, pp. 229-235, Faculty of Electrical Engineering and Computer Sciences, Maribor, Slovenia, 2002

Kačič Z., Zemljak M. (1999). SAMPA - computer readable phonetic alphabet. The WEB portal of Department of Phonetics and Linguistics, University College London. http://www.phon.ucl.ac.uk/home/sampa/slovenian.htm

Klatt D., (1987). Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, 82 (3), pp. 737-793, Sept., 1987

Kompe R., (1997). Prosody in Speech Understanding Systems, *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence*, 1st Ed., ISBN: 978-3540635802, Springer Verlag, Berlin Heidelberg,

Mihelič F., Gros J., Nöth E., Dobrišek S., Žibert J. (2000). Recognition of Selected Prosodic Events in Slovenian Speech, *Proceedings of the 2nd Conference on Language Technologies*, pp. 45-48, ISBN 961 6303-25-2, Cankarjev dom, Ljubljana, Slovenia, Oct. 17-18, 2000, Institut Jožef Stefan, Ljubljana

Müller A. F., Stergar J., Horvat B., (2002). Designing Prosodic Databases for Automatic Modeling of Slovenian Language in a Multilingual TTS System, *Proceedings of the 3rd international conference on Language resources and Evaluation, LREC 2002*, Las Palmas, Canary Island, Spain, May 18-26, 2002

Müller A. F., Tao J., Hoffmann R. (2000_a). Data-driven importance analysis of linguistic and phonetic information, *Proceedings, 6th International Conference on Spoken Language Processing* (*ICSLP 2000)*, pp. 227-230, Beijing, China, Oct. 16-20, 2000

Müller A. F., Zimmermann H.G., Neuneier R. (2000_b). Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2000), vol. 3., pp.1285-1288, Istanbul, Turkey, June 5-9, 2000

Prechelt L. (1998). Early Stopping - But When?, In: *Neural Networks: Tricks of the Trade*, Orr G. B. and Müller K. R., (Ed.), pp. 55-69, ISBN 3-540-65311-2, Springer Verlag, Berlin, 1998.

Rojc M., Kačič Z. (2000). Design of Optimal Slovenian Speech Corpus for use in the concatenative Speech Synthesis System, *Proceedings of the 2rd international conference on Language resources and Evaluation (LREC 2000)*, pp. 321-325, Athens, Greece, May 31 – June 2, 2000

Rumelhart D. E., Hinton G. E., Williams R. J., (1986). Learning internal representations by error propagation, In: *Parallel Distributed Processing: Explorations in the Microstucture of Cognition, Vol. 1, Foundations*, Rumelhart D. E. and McClelland J. L., (Ed.), pp. 318-362, The MIT Press/Bradford Books, ISBN 978-0262181204, Cambridge, MA, USA July, 1996

Sluijter A., van Heuven V., 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96)*, pp. 630-633, Wynham Franklin Plaza Hotel, Philadelphia, PA, USA, Oct. 3-6, 1996

Sonntag G. P., Portele T., Heuft B. (1997). Prosody generation with a neural network: Weighing the importance of input parameters. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, Vol. 2, pp. 931-934, Apr. 21-24, 1997

Sproat R., Olive J. (1995). An Approach to Text-to-Speech Synthesis, In: *Speech Coding and Synthesis*, Kleijn W. B., Paliwal K. K., (Ed.), Elsevier Science Inc., ISBN 978-0444821690, New York, NY, USA, December, 1995

Stergar J., Horvat B. (2003). An Environment for Word Prominence Classification in Slovenian Language, *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pp. 2087-2090, Universitat Autónoma de Barcelona, Barcelona, Spain, Avg. 3-9

Stergar J., Hozjan V. (2000). Steps towards preparation of text corpora for data driven symbolic prosody labeling, *Proceedings of the 2nd Conference on Language Technologies*, pp. 82-85, ISBN 961 6303-25-2, Cankarjev dom, Ljubljana, Slovenia, Oct. 17-18, 2000, Institut Jožef Stefan, Ljubljana

Stergar J., Hozjan V., Horvat B. (2003). Labeling of Symbolic Prosody Breaks for the Slovenian Language, *International Journal of Speech Technology*, Vol. 6, No. 3, pp. 289-299, July, 2003

Tamburini F., (2002). Automatic detection of prosodic prominence in continuous speech, *Proceedings of the 3rd international conference on Language resources and Evaluation, LREC 2002*, pp. 301-305, Las Palmas, Canary Islands, Spain, May 18-26, 2002

Traber C., (1992). F0 generation with a database of natural F0 patterns and with a neural network, In: *Talking Machines: Theories, Models and Designs*, Bailly G., Benoit C. and Sawallis T. R., (Ed.), pp. 287-304, Elsevier Science Ltd., ISBN: 978-0444891150, North-Holland

van Santen J., Olive J.(1990). The Analysis of Contextual Effects on Segmental Duration, *Computer Speech & Lanuage*, Vol. 4, Issue 4, pp. 359-390, Oct., 1990

van Santen J., Mishra T., Klabbers E., (2008). Prosodic Processing, In: *Handbook of Speech Processing*, Benesty J., Sondhi M., Huang Y. (Ed.), pp. 471-487, Springer, ISBN: 978-3-540-49125-5, Springer Berlin Heidelberg

Vereecken H., Martens J. P., Grover C., Fackrell J., Van Coile B. (1998). Automatic prosodic labeling of 6 languages, *Proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP 1998)*, Vol. 4, pp. 1399-1402, Sydney Convention and Exhibition Centre, Darling Harbour, Sydney, Australia, Nov. 30 – Dec. 4, 1998

Vereecken H., Vorstermans A., Martens J. –P. and Van Coile B. (1997). Improving the Phonetic Annotation by means of Prosodic Phrasing, *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, Vol. 1, pp. 179-182, Rhodes, Greece, Sep. 22-25, 1997

Zimmermann H. G., Neuneier R., Grothmann R., (2002). Modeling of Dynamical Systems by Error Correction Neural Networks, In: *Modeling and Forecasting Financial Data, Techniques of Nonlinear Dynamics*, Soofi A. and Cao L., [Ed.], Vol. 2, pp. 237-262, Kluwer Academic Publishers, ISBN 0792376803, Boston/Dodrecht/London

Zimmermann H. G., Müller A. F., Erdem C., Hoffmann R., (2000). Prosody Generation by Causal Retro-Causal Error Correction Neural Networks, *Workshop on Multi-Lingual Speech Communication*, pp. 116-121, Kyoto, Japan, Oct. 11-13, 2000

19

Implementing Innovative IT Solutions with Semantic Web Technologies

Vili Podgorelec and Boštjan Grašič
University of Maribor
Slovenia

1. Introduction

With modern transportation, communication, and business connections, distances are becoming narrower and competition tougher. Therefore, successful companies nowadays need to adapt to changes in environment more rapidly than they used to. Besides ever rapidly changing environment, an organizational shift towards customer has been noticed. For the last ten years or so there has been a steady international move towards changing the way customer services are delivered, financed and regulated, with the main purpose being the improvement of efficiency so that more customers could receive better service more quickly without reducing (and possibly increasing) the quality.
On the other hand, the world is witnessing a remarkable proliferation of new knowledge. With the development of information technology the amount of various business data being created and stored is growing exponentially. The endeavour of researchers and engineers to take advantage of this new knowledge, coupled with an increasing need to use limited resources more efficiently, has presented unique challenges. As new knowledge and solutions to systemic problems is sought for, it is appropriate to ask how the revolution in information and communications technologies may facilitate efficiency and prevent unnecessary duplication of effort.
Many times it has been proven that the proper use of proper knowledge is the best way of optimizing work processes. In order to improve the usability and performance the effort duplication should be minimized, access to information resources improved and possibly unified, and the information technology utilized in such a manner that would allow users to make advantage of information they need without having to bother with the abundance of them.
As the evolution of information technology and software design progresses the possible solutions to the above idea could be knowledge management and web-based software services, combined within a unified technology. Based on our experiences in developing software solutions, it is our belief that semantic web technologies could be the one technology to solve this task. In the following sections we will try to present the properties of this technology together with some advices on how to utilize it properly.

2. Semantic web technologies

The idea behind semantic web is fairly simple. Main idea is that computers would be able to understand the meaning of the data. Passin defines semantic web as a vision or a liquid,

developing and informally defined concept (Passin, 2004). According to Passin, vision of semantic web is that computers would be able to find, read and understand the meaning of data. Tim Berners-Lee, sees semantic web as "web of data" compared to web of documents as we know world wide web today (Berners-Lee et al., 2001).

There exist many scenarios of semantic web usage. Most of them include autonomous agents that are able to autonomously find data on the web of data and present information that is relevant to us. If we look at the situation from today's point of view, one would have to search the search engines about the key-word of the subject of interest, then he would have to search the resulting web pages for information, he is interested in.

Technologically speaking we are still far from the discussed scenario, but there already exist technologies that should enable machines to operate with data and its meaning. These technologies are called semantic web technologies (SWT). There were many attempts to build core SWT – according to Passin, the most promising are technologies that are supervised by W3C consortium (Passin, 2004). These technologies are mostly built upon technologies from the Defense Advanced Research Projects Agency (DARPA) and their DAML (DARPA Agent Markup Language) language. SWT are based on XML language that enables them to be platform and program language independent. SWT are extensible like XML. This means that for each purpose there exists a separate technology that is compatible with others.

SWT are built in layers, as shown on Figure 1. Each upper layer provides additional functional aspect and is based on the lower one, with which is fully compatible. Most bottom layer includes Unicode and URI technologies. Unicode enables SWT to be platform and language independent. URI is not just used for electronic resource identification, but for general resource identification. Each resource, also a physical one, is in the semantic web represented with an URI descriptor.

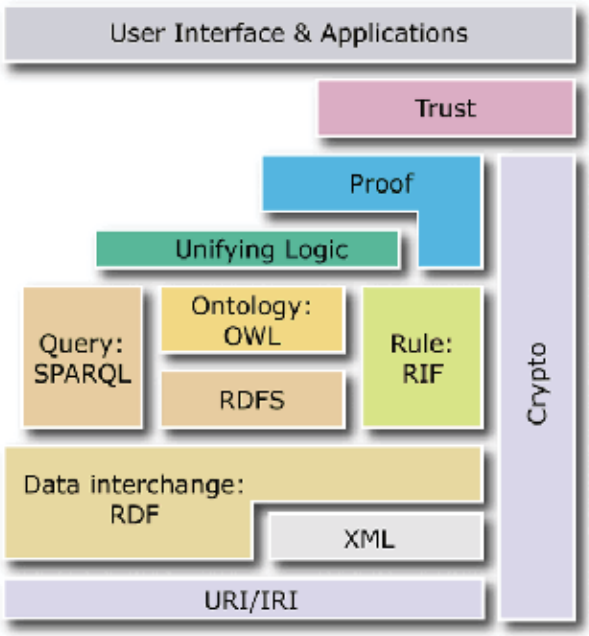


Fig. 1. The layers of semantic web technologies

The second layer of enabling technologies for SWT includes XML, XML Schema and namespaces. Semantic web data is represented in XML language, this implicates that semantic web documents are actually XML documents with predefined schema.

2.1 Describing resources within semantic web technologies: RDF and ontologies

Above the enabling technologies with which we can get only self describable documents is RDF layer. RDF is core SWT and is acronym for resource description framework. It is used for describing resources of any kind. Resources are represented with an URI descriptor. RDF document is a composition of statements called triples. Each triple is a statement about a resource and is composed of subject, predicate and object. While subjects and predicates are URI resources, an object can be either an URI resource or a plain string called literal. Literals cannot be identified; therefore they cannot be referenced and used as objects or predicates (Lassila & Swick, 1999).

Because of RDF's flexibility and its simple structure, it can be used to describe practically any resource. Because of its structure, an RDF document can be represented as a directed graph, where subjects and objects are represented as nodes, while predicates are represented as edges.

Above the RDF layer in SWT stack is ontology. Ontology is a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations (Wordnet[1]). For ontology description there exist more languages, most common nowadays is OWL (Web Ontology Language), which is derived from DAML and OIL. OWL is vocabulary extension of RDF, which is why an OWL document is also a valid RDF document. Rules and characteristics of RDF also apply to OWL (Passin, 2004).

OWL uses description logic to classify resources defined in a RDF document. OWL 1.1 is divided in three subsets: OWL Full, OWL DL and OWL Lite. Instead of using these sublanguages, OWL 2 introduces profiles (OWL 2 EL, OWL 2 QL, OWL 2 RL). Difference between different sublanguages is in number of language constructs they support. Different sub-sets were developed because of compensation between performance and expressiveness. OWL Lite has the best performance and limited expressiveness, while OWL Full has opposite characteristics. DL in OWL DL stands for description logic. It was designed for existing description logic business segment and should be preferred over OWL Full in production environments. Main advantage over the full subset is that OWL DL is fully decidable (Dean & Schreiber, 2004).

We already mentioned that OWL is used for classification purposes. It has two main categories: classes and properties. RDF nodes (subjects and objects) are categorized in OWL classes, while RDF predicates are categorized by OWL properties. Nodes and predicates are being classified using description logic predicates. Classes in OWL are treated as sets and OWL supports all set operations (union, intersection, disjoint, equivalent, subset) (Dean & Schreiber, 2004).

Next layer which is also important in SWT stack is logic layer. Rule based languages are natural choice for this purpose. W3C consortium has released a member submission called SWRL (Semantic Web Rule Language). SWRL is rule-based language that is combination of OWL Lite and OWL DL subsets with Datalog RuleML sublanguages. With SWRL language, knowledge engineer can define rules that are then applied to OWL knowledge database.

[1] http://wordnet.princeton.edu

Rules are composed of antecedent and consequent. If the requirement defined in antecedent is met then the triple in consequent is inserted in the graph. The two top most technologies (trust and proof) in SWT stack are not supported yet. Because there is no technical solution to these problem areas available we will ignore these two layers.

2.2 The key role of ontology

What is the purpose of ontology in semantic web? Ontology describes the subject domain using notions of concepts, instances, attributes, relations and axioms. Ontology can be defined as a formal explicit specification of a shared conceptualization. It is a useful way to organize and share information while offering intelligent means for knowledge management. Ontology also enhances semantic search in distributed and heterogeneous information services. Ontologies are the key player, if we want to do (automatic) search in more advanced ways, not only keyword search.

There are several benefits of using ontologies for information solutions. Semantic search engines return instances that constitute answers to queries rather than documents containing search strings as in keyword search engines. Semantic search uses meanings (semantics) of the query terms defined in the ontology. The data of ontology constitutes precise answers to user questions. Users can further browse related concept because answers are interconnected through semantics. It can be speculated that using ontology supported systems users will also be able to invoke functionalities or query data using free text input in the future.

The central part of a semantic web application is an ontology that describes some knowledge domain using notions of concepts, instances, attributes, relations and axioms. It is a useful way to organize and share information while offering means for enhanced semantic search in distributed and heterogeneous information systems. Ontology can be defined as a formal explicit specification of a shared conceptualization.

Nowadays, there are many application domains, where the utility of ontologies is widely accepted, and where ontologies have already been deployed at large scale. However available ontologies, actually used as common vocabulary for certain applications, cover particular domains to different granularities and cannot be directly used for the semantic web. This issue has been addressed for example in the medical domain in project GALEN[2], where the authors developed a special representation language, tailored for the particularities of the (English) medical vocabulary. However, the usage of a proprietary representation makes the ontological knowledge difficult to be extended by third parties or in a semantic web setting.

In order to adopt the SWT for supporting the information management in a specific domain, there is a key field that needs to be addressed: domain knowledge that we want to integrate within the domain. For this purpose an ontology needs to be defined, which will then allow all further actions, like semantic annotation of data (in accordance with the ontology), integration of data resources, advanced searching and inferring on the data.

It should be pointed out, that there are two notions of the term ontology: heavyweight and lightweight ontology. Heavyweight ontologies are mainly used for complex, logic based modeling of a knowledge intensive domain (e.g. gene ontology, protein ontology). They are carefully designed throughout a strenuous and vigorous process involving a consortium of

[2] http://www.opengalen.org

experts; the specification of concepts, relations and logical restrictions is very precise. On the other hand, lightweight ontologies are used mainly for data integration purposes or they act as a common vocabulary; in this way concepts may be defined more loosely and are practical goal oriented. In this paper we are using the term ontology as a lightweight ontology used for integrating information resources.
The integration of information (or knowledge) within a specific domain has always been a significant issue. The introduction of ontologies and SWT have provided some promises in this direction. Some knowledge integration approaches using SWT have already shown some results, either as an integrative mechanism within an organization (Lenz et al., 2007) or inter-organizationally (Knaup et al., 2007).

3. Common SWT based system architecture

As already said, SWT are usually represented in a layer cake model (Figure 1), where XML and URI are foundations for SWT building blocks. First SWT specific building block is RDF (Manola & Miller, 2004) – a language for describing resources, which are represented as URI-s. Next important building block is OWL (Dean & Schreiber, 2004) – a language for describing ontologies. Next to ontology building block is RIF, which stands for rule interchange format and includes rule languages that are compatible to RIF. One such rule language widely used in SWT is SWRL – semantic web rule language. It enables reasoning on concepts defined in OWL ontologies.
A typical SWT system is based upon RDF, OWL and a rule language compatible to RIF (SWRL is widely used). In this manner, RDF is mainly considered as a data backend and a data interchange technology. Concepts that are defined in ontology are mostly used for data integration. Rules enable encapsulation of business logic; based on defined rules, new knowledge is being inferred according to concepts defined in ontology and RDF data. A query language (like SparQL) is used to query the semantic data. These building blocks represent the core SWT. A typical SWT based system architecture is shown in Figure 2.

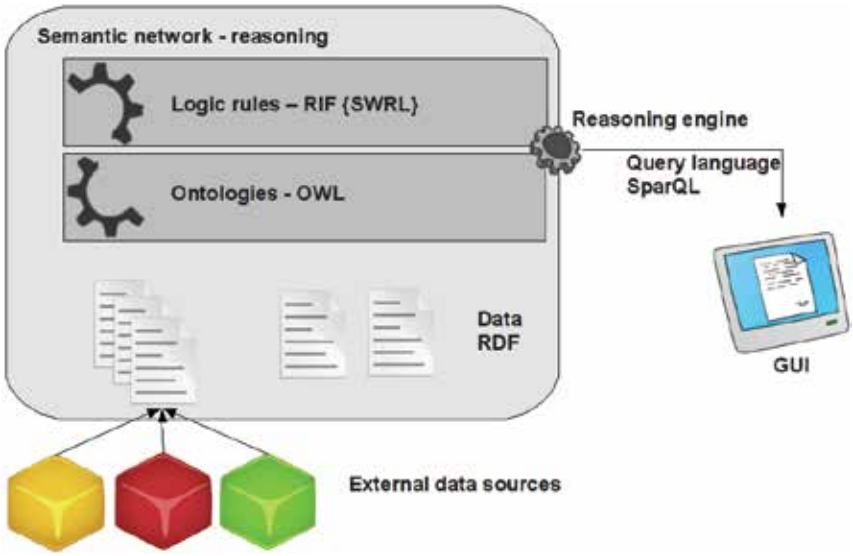


Fig. 2. Common SWT based system architecture

SWT provide means for knowledge representation. Opposite to past research efforts in the field of artificial intelligence, where knowledge representation systems were mostly centralized and isolated, SWT are designed to be used on a Web scale in heterogeneous information environment. While SWT provide fairly expressive formalisms for knowledge representation and reasoning, this often results in poor performance.
To overcome this issue, we propose division of information space represented in semantic network into two subsets:

- static information – information that is essential for reasoning purposes, hence has to be always available to the reasoner,
- dynamic information – information that is not needed for reasoning purposes, thus it can be fetched when needed.

By dividing information space into two subsets and excluding one from semantic network, the size of the semantic network is reduced; hence the amount of information that needs to be processed by the reasoner is reduced. For providing dynamic information, we propose use of semantic web services that are executed automatically, when the demand for dynamic information arises.
This way not only the performance is increased, but the interoperability with legacy systems is increased also. Data provided by services is not limited to data that is excluded because of performance issues. Based on service oriented architecture paradigm, services can also act as information providers from legacy systems in loose coupling, higher reuse and greater interoperability manner.
Having in mind two mayor drawbacks of a common SWT based architecture, namely poor performance on large data sets and high data integration and integrity costs with external data sources, instead of using agents to import RDF data to semantic network, web services are used.
In this manner, the new SWT based system framework substitutes some data sources with web services that provide data on demand. This way, dynamic data as defined earlier, is being provided by web services. To be able to combine static and dynamic data in a transparent manner, the framework has to be able to automatically execute Web Services, when the need for data they provide arises. To solve the above problem we propose to incorporate SOA (Service oriented architecture) principles (Erl, 2005).

4. Incorporating SOA concepts

Though SWT are very suitable for integration purposes, we identified some short-comings in this approach. SWT are very flexible and we can transform practically any machine readable data into RDF and then process it based on the ontology. To be able to integrate data, we have to import it into the semantic network first. When the intention is not to reason on integrated data, but rather just to integrate it, there are other more suitable methods for data integration. Because of the fact, that we have to import the data into the semantic network to be able to integrate it, there arise two main problems: (1) the size of the semantic network can grow at a very fast rate, because of this there can be performance issues and (2) by importing data into the semantic network, we basically replicate the data. In order to have accurate information, we have to synchronize the data between the originating data source and the semantic network.
In a SWT system, we can have information, that is crucial for the reasoning process and also information that is not used for the reasoning process. This kind of information is used for

providing additional information to the user. For example, suppose we want to support medical decision making process of selecting the most suitable physician for a support request. To be able to reason about the physician's competences, we need information about his previous experiences and education. To be able to propose a suitable physician, we need information about the nature of the problem and competences of all the physicians.
Let us suppose that this information is sufficient for successful physician selection. To be able to propose a physician, we have to import this information into semantic network. While this information is sufficient for the reasoning process, the user, that uses knowledge management system may have needs for additional information like contact information of the patient requesting physician, previous records for this particular patient etc. Usually this information is already stored in some other information system and synchronization can result in high development costs and larger amount of processing.
Therefore we propose use of services to provide information that is not essential to reasoning process. Information space can be divided into two subsets: static and dynamic data. Static data is data that is needed for the reasoning process and should be always available in the semantic network. Dynamic data, on the other hand, is not needed by the reasoning process – it can be provided dynamically when the need for it arises. In this manner, we propose the use of services to provide these dynamic data (Figure 3). That way we don't need to import a large amount of data into the semantic network. Instead, the data is being requested when the need for it arises. If we return to our example, when the knowledge system proposes a particular physician, if the user has need to contact the patient (physician requester), the service that provides patient contact information can be automatically executed and the user can be seamlessly presented with contact information.
In a SWT system the data is described semantically. This means that computers have an awareness of the meaning of data. In the same manner we can semantically describe services. If we model and integrate semantics of the data and services in an ontology, the services can be executed automatically. That way information can be presented to the user transparently in both cases, whether it is stored in the semantic network or provided by a service. Relations between data semantics and its role in organization knowledge can be modelled in an ontology.

4.1 The principles and benefits of SOA

Service oriented architecture (SOA) is an evolutionary step from enterprise application integration (EAI). Main purpose and goal of both EAI and SOA is integration of information between several different information systems. EAI creates tight connections between different information systems in such way, that each information system imports data from other information systems. While this is fairly natural way of integration of information, it is often connected with high maintenance efforts and low level of reuse and flexibility.
Opposite to tightly coupled EAI, SOA introduces loosely coupled, distributed and flexible architecture. SOA introduces concepts of services. While EAI creates connections between each pair of applications, applications in SOA environment expose their data and also functionality in form of platform and program language independent services. Applications that need information or other applications functionality only have to query the service. By also exposing functionality greater level of reuse can be achieved.
According to (Erl, 2005), the key aspects of SOA are: (1) loose coupling – services maintain a relationship that minimizes dependencies and only requires that they retain an awareness of

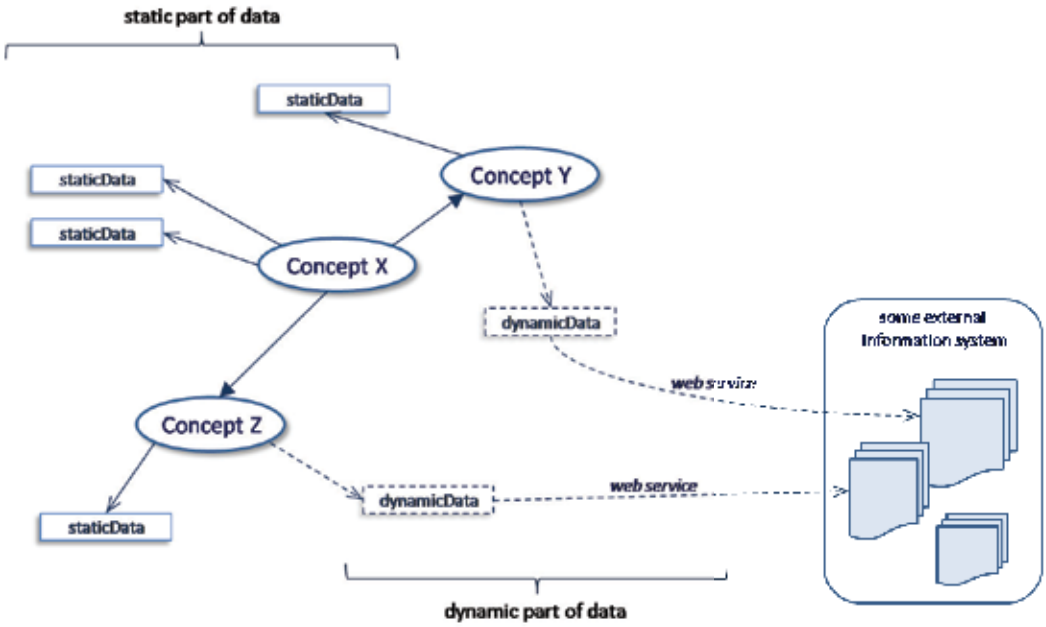


Fig. 3. Basic concept of decoupling semantic information into static and dynamic (provided by an external information system) data

each other; (2) service contract – a communications agreement, as defined collectively by one or more service descriptions and related documents; (3) autonomy – services have control over the logic they encapsulate; (4) abstraction – beyond what is described in the service contract, services hide logic from the outside world; (5) reusability – logic is divided into services with the intention of promoting reuse; (6) composability – collections of services can be coordinated and assembled to form composite services; (7) statelessness – services minimize retaining information specific to an activity, (8) discoverability – services are designed to be outwardly descriptive so that they can be found and assessed via available discovery mechanisms.

SOA is basically a concept and is independent of the implementation technologies. When the term service oriented architecture is used, it is mostly meant in a context of web services. In this manner, (Erl, 2005) defines contemporary SOA as SOA, that is built around web service technologies. When we use the term SOA in this paper, we actually mean contemporary SOA as defined by (Erl, 2005).

The architecture, presented in the following sections, can be seen as an extension of SOA. The same service can be reused for automated ontology based data integration, as described in this paper, as well as in other common SOA aware applications.

4.2 Semantic web services execution ontology (SWSEO)

Web services (WS) technology mainly provides standards for functional and also nonfunctional description, e.g. quality of service, security, authorization. While they provide very good technical platform, they lack in providing semantics to the services. This results in worse than expected discovery, reuse and composition of web services. To overcome these issues, semantic web services (Akkiraju, 2006) were introduced.

Semantic web services (SWS) combine concepts from semantic web and web services. As Grosof (2003) pointed out, semantic web services can be understood in two ways: as (1) semantic [web services] and as (2) [semantic web] services. The former concept relies more on WS and is used for knowledge-based service descriptions (discovery, execution, composition of WS), while the latter concentrates more on SW concepts and is used mainly for knowledge and information integration. In our framework we use SWS in both contexts, as S[WS] for capturing IT support knowledge and as [SW]S for supporting representation of operational knowledge (classic KMS).

There are several approaches for implementing SWS, the most common are: WSMO, OWL-S and SAWSDL. There are slight differences in basic concepts. We chose to use SAWSDL because of the following facts: (1) we don't need complex discovery and composition capabilities in our framework; thus we can use a simpler formalism, (2) WSMO uses its own language that is not compatible with SWT, (3) research and development effort of OWL-S is fading and a lot of tools are already outdated, (4) existing WS can be easily converted to SAWSDL by just semantically annotating WSDL, (5) SAWSDL is not just compatible with SWT but it is also compatible with current Web Services, (6) SAWSDL is interoperable with SOA implementations.

SAWSDL does not specify how services are modeled. Because of that, we have developed a lightweight service modeling ontology that is targeted at automated execution of web services. The ontology is called semantic web services execution ontology (SWSEO). It is defined in OWL-DL and is shown in Figure 4.

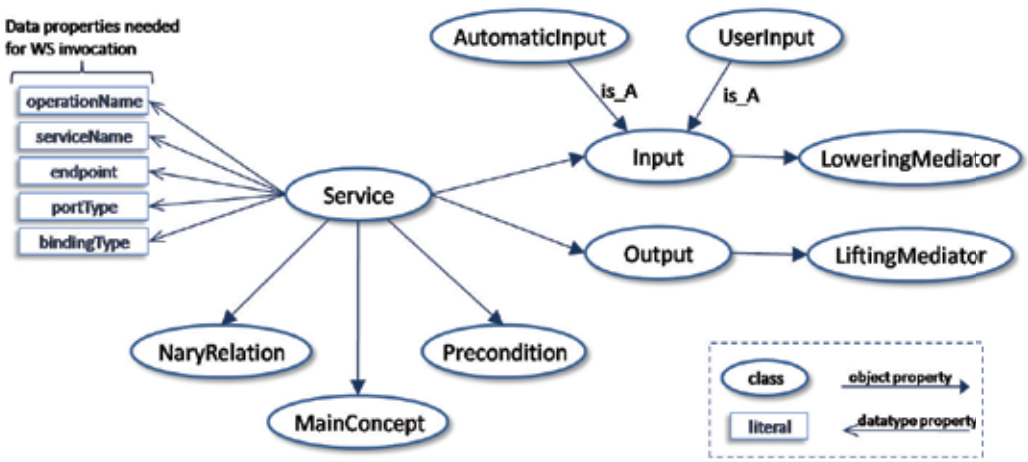


Fig. 4. Basic concepts of Semantic Web Services Execution Ontology (SWSEO)

Main concept in SWSEO is service. Opposite to WSDL, SWSEO service is actually a WSDL operation. It has five data properties (presented on the left hand side in Figure 4) which are used for web service invocation and are parsed from SAWSDL document at the registration time.

Every service has also some input and output messages. Web Services are based on XML language and use XML documents as input and output messages to services. To be able to use this data in semantic networks; XML data has to be converted to RDF format and vice versa. In case of input messages, semantic data represented in RDF has to be lowered to XML documents and in case of output messages XML data has to be lifted to RDF format.

For this purpose we defined lifting and lowering mediators that take care of data conversion.
Lifting mediators can be XSL transformations (Kay, 2007), Java classes or even other Web Services. Lowering mediators on the other hand have one additional mediator type: SparQL mediator. In RDF the same data can be represented in different ways. Because of that, XSLT, which does transformation based on XML document structure, cannot handle all different data representation variations. For this purpose a combination of SparQL query and XSLT can be used. SparQL takes care of getting the right structure, after that XSLT is used to actually lower the data. As seen in Figure 3, input messages can be of two types: (1) automatic input – data is being fetched automatically from the semantic network by the framework and (2) user input – input data is not stored in the semantic network, but it is provided by the user when querying semantic network.
There are three further concepts, we haven't mentioned yet. First is precondition. This concept is used in case more than one service provides the same type of data and the service that is being invoked is selected based on the instance, e.g. let us suppose we have an e-health application that uses web services for getting some medical equipment availability information.
Each participating equipment provider provides its own availability service; all the availability services provide same type of data (equipment availability). Service that is being invoked is selected based on precondition. Precondition defines for which participating provider a particular service provides medical equipment availability information.
Concept named main concept is used for extracting automatic input messages from semantic network and for creating service execution plan. For example described in previous paragraph, main concept would be participating medical equipment provider. This means that the framework has to check equipment availability by executing the web service for each given provider.
Last basic concept of SWSEO is nary relation. RDF and OWL support only binary relations between concepts. If we want to make a nary relation, we have to create a holding class (Hayes & Welty, 2006). There are situations where services provide nary relations for which holder class instances are not yet in the semantic network. This class actually serves as an instruction for the framework to create the holding class instance and connect it with the main concept.

4.3 Automated semantic web services execution architecture

Figure 5 shows system architecture for automated execution of SWS in accordance with SWSEO. The system acts like a wrapper to the SparQL endpoint. The whole process of service input retrieval, data conversion, service execution and model integration is transparent to the user; the user has to provide only the SparQL query.
Main components of the architecture are: (1) SparQL query processor – it splits a query into static and dynamic part, returns concepts from query that are service outputs and parses user service input information from where clauses, (2) input provider – provides data defined as automatic input, (3) input and output mediators – provide lifting and lowering capabilities (Java, XSLT, and SparQL mediators), (4) service executor – invokes web services, (5) query executor wrapper – provides data from semantic networks needed by other components in an efficient way by caching data and reducing the number of query executions.

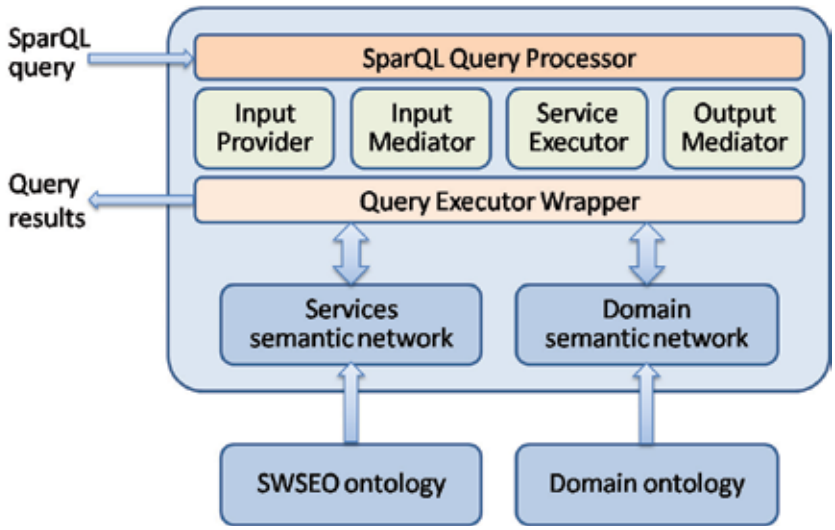


Fig. 5. Automated Semantic Web Services execution architecture

5. Improved SWT based system architecture

Based on all the concepts discussed above, we can now finally represent the improved system architecture framework based on SWT that should be a possible solution to the idea of the unified information system architecture for software systems proposed in the introduction of this chapter. The conceptual system architecture is presented on Figure 6.

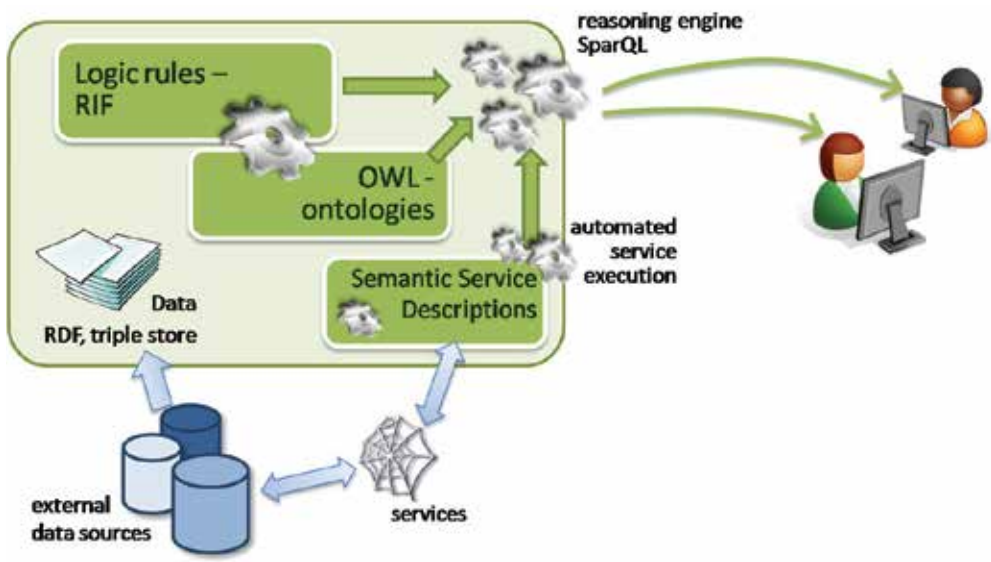


Fig. 6. Advanced SWT based system architecture

The presented system can collect data from both static data repositories (relational databases, intranet site, file systems, etc) and also from various web services. In this manner, the integration with other information systems should be easier to achieve, especially when

their data have already been exposed in a form of web services. By exposing data as web services also the size of the semantic network can be reduced, which results in better performance. In order to be able to fully use the advantages of semantic based system, the services should be described with a proper semantic service description technology (as discussed previously).
The proposed architecture is driven by a set of specific ontologies for different views within the whole medical process, which should be interconnected in order to fully explore the possibilities of semantic data. The system is open to connect to various publicly available semantic databases, if it is appropriate for a user. The processing of the data in the system is by no means limited to a single technology. All known technologies, tools and methods for efficient processing of semantic data can be used.
The reasoning engine that provides the data to users over different user interfaces (web or desktop applications, data repositories endpoints . . .) is defined on a set of standardized technologies, such as logic rules (RIF), SparQL query language, etc. In this manner, the final processing system can be defined in such a way that best suits the software requirements of a specific implementation.

6. A case study: project team building

To test the appropriateness of the presented improved SWT based system architecture for implementing an innovative IT solution we decided to develop a system for supporting the building of project teams regarding the requirements of a project and skills of potential team workers. For a technical project, let's say in software engineering, to be successfully implemented there is a need for bright, skilled individuals with good technical skills and exceptional attention to detail. However, the real world projects nowadays normally require more than just good individuals. Even a group of great individuals is not enough. What we really need is a team – a team of cleverly selected individuals, who will combine their personal technical skills with their teamwork skills in order to achieve the project goals. What we need to compose great teams is a proper team building approach and a supporting technology to implement the approach. It is our belief that the improved SWT based system architecture, presented in previous sections, is a good and valid approach to project teams building.

6.1 Team building overview

Team building is an effort in which a team studies its own process of working together and acts to create a climate that encourages and values the contributions of team members. Their energies are directed toward problem solving, task effectiveness, and maximizing the use of all members' resources to achieve the team's purpose. Sound team building recognizes that it is not possible to fully separate one's performance from those of others. Team building works best when the following conditions are met (Frances & Young, 1979):

- There is a high level of interdependence among team members. The team is working on important tasks in which each team member has a commitment and teamwork is critical for achieving the desired results.
- The team leader has good people skills, is committed to developing a team approach, and allocates time to team-building activities. Team management is seen as a shared function, and team members are given the opportunity to exercise leadership when their experiences and skills are appropriate to the needs of the team.

- Each team member is capable and willing to contribute information, skills, and experiences that provide an appropriate mix for achieving the team's purpose.
- The team develops a climate in which people feel relaxed and are able to be direct and open in their communications.
- Team members develop a mutual trust for each other and believe that other team members have skills and capabilities to contribute to the team.
- Both the team and individual members are prepared to take risks and are allowed to develop their abilities and skills.
- The team is clear about its important goals and establishes performance targets that cause stretching but are achievable.
- Team member roles are defined, and effective ways to solve problems and communicate are developed and supported by all team members.
- Team members know how to examine team and individual errors and weaknesses without making personal attacks, which enables the group to learn from its experiences.
- Team efforts are devoted to the achievement of results, and team performance is frequently evaluated to see where improvements can be made.
- The team has the capacity to create new ideas through group interaction and the influence of outside people. Good ideas are followed up, and people are rewarded for innovative risk taking.
- Each member of the team knows that he or she can influence the team agenda. There is a feeling of trust and equal influence among team members that facilitates open and honest communication.

Team building will occur more easily when all team members work jointly on a task of mutual importance. This allows each member to provide their technical knowledge and skills in helping to solve the problem, complete the project, and develop new programs. During this process, team building can be facilitated as members evaluate their working relationship as a team and then develop and articulate guidelines that will lead to increased productivity and team member cooperation. Team performance can best be evaluated if the team develops a model of excellence against which to measure its performance.

6.2 SWT based team building approach

We decided to develop the whole system in a manner of a semantic web portal, which serves as an entry point to our solution in project team building. It supports users in building project teams effectively and efficiently. The architecture of the portal is presented in Figure 7.

This architecture provides a mean to manage both members' and projects' profiles through a web server by members themselves, by project leaders and by project administrators. The inference engine uses the profiles together with previous projects' data in order to propose members for any new project regarding the requirements. The system's inferring capabilities can be improved by managing the skills matching database which is used to reveal the hidden skills of members, not provided directly by them or the project leaders.

During the project cycle the portal should not be used too frequently. If we would push users to use it over and over again, we would disturb project activities and waste team's energy. However, we want results from the portal – so we designed it to be used only on beginning and ending of a project (Figure 8). If project team uses information support for their project activities, portal should integrate their existing data, so no changes in work are necessary. When using some groupware to support project work, the data could be gathered automatically.

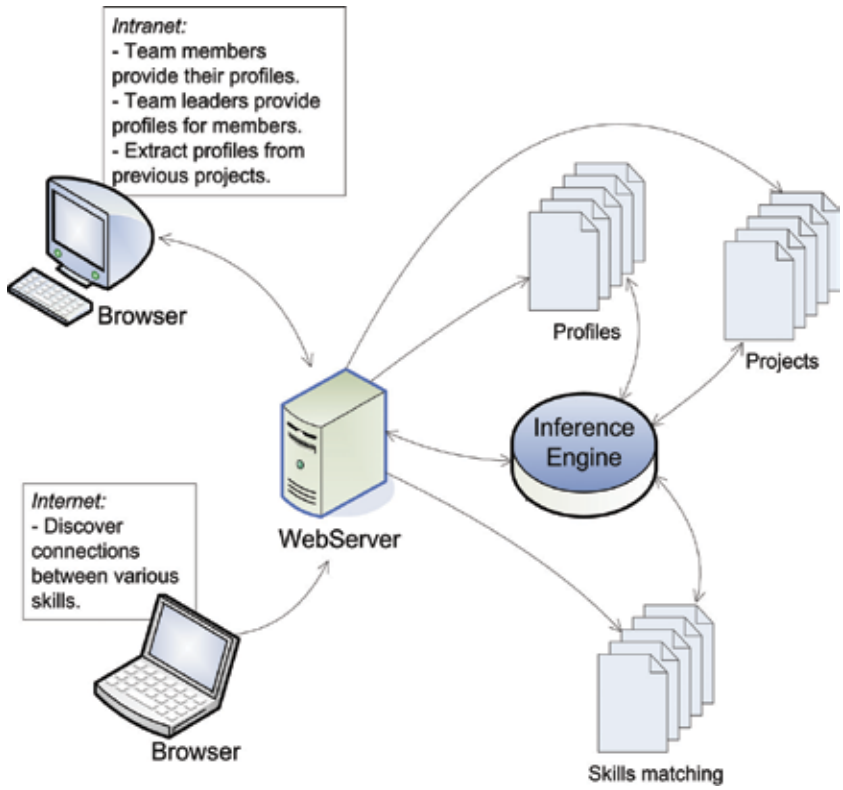


Fig. 7. The architecture of a semantic web portal for project team building

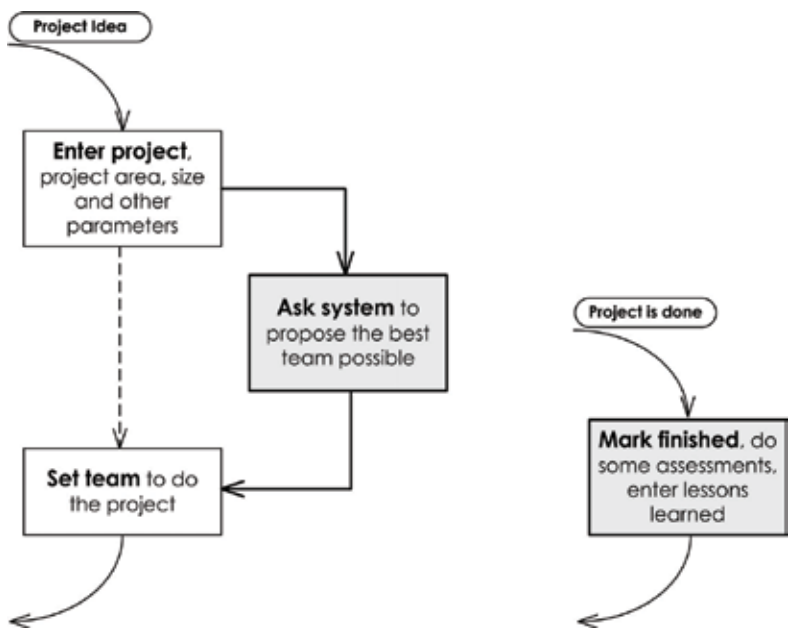


Fig. 8. The role of portal for project team building during project cycle

6.3 Ontology-based personal skill management

An overview of the related work in ontology-based personal skill management is presented in Biesalski & Abecker (2005). Already Stader & Macintosh (1999) and Jarvis et al. (1999) promoted the idea of ontology-based modelling of personnel skills and job requirements – as part of comprehensive, workflow-oriented enterprise modelling. There, the following potential applications of ontology-based skill profiles are listed:

- skill gap analysis – at the enterprise level, as a part of strategic HR planning,
- project team building,
- recruitment planning – again a part of strategic HR planning,
- training analysis – at the level of individual personnel development.

Those approaches were mainly technology-driven and were – to our knowledge – never realized in a large-scale industrial environment. Nor have they been accepted by the HRM departments, translated into HRM people's terminology, embedded into more comprehensive models and procedures of HRM people, and integrated with existing software infrastructures. After those first publications, there were a number of interesting technology-oriented researches which showed that in particular skill matching can benefit from interesting technological approaches, such as background knowledge exploitation. For instance, Liao et al. (1999) employs declarative retrieval heuristics for traversing ontology structures. Sure et al. (2000) derives competency statements through F-Logic reasoning and developed a soft matching approach for skill profile matching. Colucci (Colucci et al. (2003)) use description logic inferences to take into account background knowledge as well as incomplete knowledge when matching profiles.

6.4 Application ontology: personal skills and project team

There have been many approaches to describe personal skills within an ontology. They included both technical skills (like knowledge of programming languages, development methods, specific tools) and inter-personal skills (like communication skills, affableness, teamwork). Based on the team building theory and our own experiences from performed projects the main items that have to be included in such ontology should be:

- formal and informal education,
- experiences,
- practical skills,
- performed projects,
- preferred tasks,
- preferred role within a team,
- communication skills,
- teamwork spirit.

As far as we could find within the recent literature, there have been some approaches to describe project teams with an ontology. However, they have not been used for project teams building. For this purpose the existing project team ontologies, which include basic information about team members, resources, purpose of the project, etc., should be extended with the following information:

- the priorities of the project,
- the importance of specific phases (regarding the current stage),
- complexity of the project (regarding the previous ones),
- the technical type of the project (prototype, research, production, etc),
- special knowledge requirements,

- preferred personnel,
- the type of the product/service being developed,
- the relation with other (especially previously successfully performed) projects.

The resulting ontology is presented in Figure 9 – only classes and object properties are shown. Ontology is used in the portal to help reasoner use metadata and construct the proper team for performing the project.

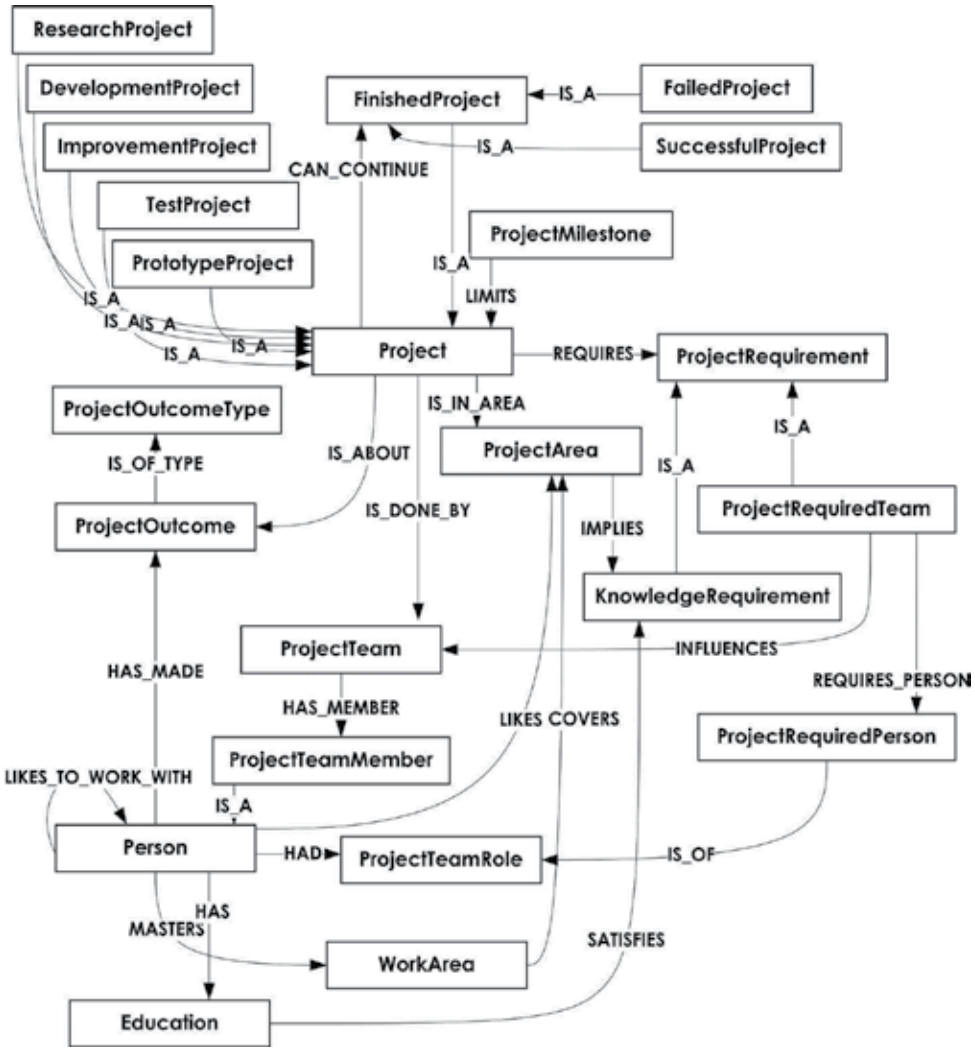


Fig. 9. The used ontology for project team members selection

6.5 Inferring on the application ontology

The project team building system prototype is implemented mainly in Java programming language using open source Jena semantic web development library[3]. It provides us with a

[3] http://jena.sourceforge.net

straight-forward development system, very appropriate for semantic web enabled applications. For the inferring part, we have used Pellet[4]. Pellet is an OWL reasoner for Java that enables monotonic reasoning on rules specified in semantic web rule language (SWRL). Using rules on top of a formal ontology enables powerful inferring capabilities.
To demonstrate rule based inference, example pair of rules are shown in Figure 10. These two rules select suitable persons for a project team based on project requirements. Concepts that are used in the rules are defined in the ontology above (Figure 9). In the example we can see that project team members can be selected based on the project requirements in two ways:

- using the information about education of a person (knowledge requirements that are covered by some education), and
- using the information about work areas mastered by a person (a work area covers several project areas and a project area implies several knowledge requirements).

$$IS\_DONE\_BY(?project, ?projectTeam) \wedge REQUIRES(?project, ?knowledgeReq) \wedge HAS(?person, ?education) \wedge SATISFIES(?education, ?knowledgeReq) \rightarrow HAS\_MEMBER(?projectTeam, ?person)$$

$$IS\_DONE\_BY(?project, ?projectTeam) \wedge REQUIRES(?project, ?knowledgeReq) \wedge MASTERS(?person, ?workArea) \wedge COVERS(?workArea, ?projectAres) \wedge SATISFIES(?projectArea, ?knowledgeReq) \rightarrow HAS\_MEMBER(?projectTeam, ?person)$$

Fig. 10. An example set of SWRL rules for inferring on the ontology.

In this way a team member can be automatically selected based on the requirements of the project. Note that an IS_A relation between knowledge requirement and project requirement has been implicitly used in this example. The knowledge base for inferring on SWRL rules is induced directly from the ontology and/or the corresponding database.
In order to create a holistic application for managing projects and team members, the application has to provide complete information about project members and projects as well. Unfortunately, this kind of information is usually stored in legacy applications. It is not reasonable to replicate all this data in semantic application. For this reason, the prototype uses services to obtain information from other applications in the manner of service oriented architecture.
Semantics of the data that is returned by Web Services is modelled in the application ontology. SWSEO based service annotations enable reasoning system to obtain data dynamically by automatically executing Web Services. Besides proposing appropriate team members as described earlier, SWSEO compatible services on the other hand enable us to dynamically integrate detail information about project members from other systems (e.g. personal information, detailed project description as well as work items, human resources availability, team member's assignments, passed certifications, etc.)

6.6 Technologies used

As shown in Figure 11 the system itself does not consist of many different technologies, which is in our belief good. As mentioned before, fundamentals for the system lies in J2EE

[4] http://clarkparsia.com/pellet

platform, XML enabled database (we used the Oracle 10gR2 database), and connection with the external web services.

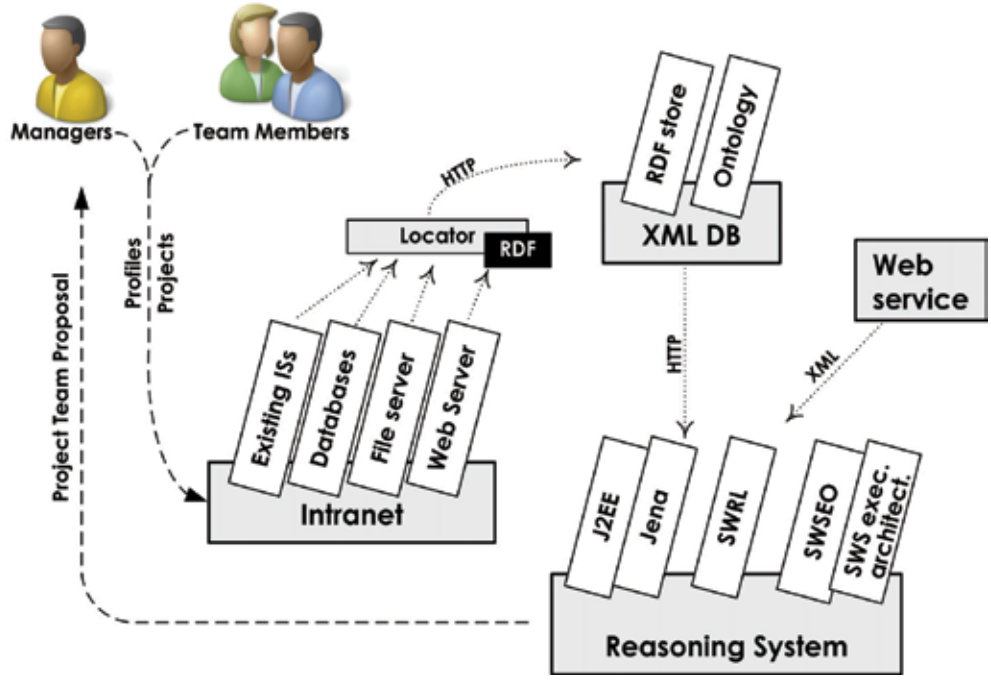


Fig. 11. The main technologies used to develop the project team building portal

First vital component, called "locator", is responsible for collecting as many internal data in RDF as possible. It extracts data from existing systems, web pages, databases and file systems. Collected RDFs and presented ontology are persistently stored in XML database and represent the static part of the whole semantic data network. The dynamic part of the data is served on demand by the web service. The both parts together are prepared for the reasoning system, based on J2EE and Jena with the use of SWRL. So the second vital component is the reasoning system, which uses the SWSEO-enabled web services, automated semantic web services execution architecture and SWRL rules to infer on the integrated semantic data.

7. Conclusions

This paper provides some results of our endeavour in adopting SWT for implementing innovative IT solutions. Performing some experiments using SWT as an enabling technology, we came to the conclusion that the common SWT based system architecture has some important drawbacks. In this manner, knowing the great potential of SWT, our goal was to find out whether improved settings of SWT can be able to overcome at least some of the difficulties encountered.

The proposed approach to the utilization of state of the art software technologies for the development of innovative IT solutions using SWT serves the purpose. In some pilot implementations the proposed improved system architecture enabled us to integrate the data from different processes at the ontology level. Furthermore, it enabled us to

interconnect our own ontology with the existing ontologies, with both semantic and relational data repositories, and also with the dynamic data from web services. Using the division of semantic data between static RDF based data repositories and external semantic web services, we succeeded to somewhat reduce the poor performance of known semantic applications. It must be said, however, that only a very limited data resources have been used in our experiments, which show no exact proof of how the system would perform when scaled to a real world software system. By exposing data as web services, the size of semantic network can be reduced, which results in better performance. Yet this does not have a deterministic behaviour and further research work should be done on this topic. On the other hand by enabling integration and automatic execution of web services better interoperability with other information systems can be achieved.

It is our intention to further improve the proposed SWT based system architecture in the future. Primarily, we would like to perform some tests on scaling properties of the proposed system architecture. Furthermore, the integration possibilities with the existing information systems, both monolithic and service based, should be evaluated in a more detailed manner. Finally, we would like to test the system in a real world environment using a defined methodology in order to evaluate whether the technology is appropriate to be used by professional software developers.

8. References

Akkiraju, R. (2006). *Semantic Web Service - Theory, Tools, and Applications*, Idea Group, chapter Semantic Web Services, pp. 191–216.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web: Scientific american, *Scientific American* .
URL: http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Biesalski, E. & Abecker, A. (2005). Integrated processes and tools for personnel development, *1th International Conference on Concurrent Enterprising, University BW Munich, Germany, 20-22 June 2005.*

Colucci, S., Noia, T. D., Sciascio, E. D., Donini, F. M., Mongiello, M. & Mottola, M. (2003). A formal approach to ontology-based semantic match of skills descriptions, *Journal of universal computer science, Special issue on skills management* 9: 1437–1454.

Dean, M. & Schreiber, G. (2004). OWL web ontology language reference, *W3C recommendation*, W3C.

Erl, T. (2005). *SOA Principles of Service Design*, Prentince Hall/PearsonPTR.

Frances, D. & Young, D. (1979). *Improving Group Work: A Practical Manual for Team Building*, University Associates, Inc.

Grosof, B. (2003). Semantic web services: Obstacles and attractions, *Panel at 12th Intl. Conference on WWW*.

Hayes, P. & Welty, C. (2006). Defining n-ary relations on the semantic web, W3C Working Group Note.

Jarvis, P., Stader, J., Macintosh, A., Moore, J. P. & Chung, P. W. H. (1999). What right do you have to do that?-infusing adaptive workflow technology with knowledge about the organisational and autority context of a task., *ICEIS*, pp. 240–247.

Kay, M. (2007). Xsl transformations (xslt) version 2.0 (w3c recommendation 23 january 2007), *Technical report*, W3C.
URL: http://www.w3.org/TR/xslt20/

Knaup, P., Garde, S. & Haux, R. (2007). Systematic planning of patient records for cooperative care and multicenter research, *International Journal of Medical Informatics* 76(2-3): 109 – 117. Connecting Medical Informatics and Bio-Informatics - MIE 2005.

Lenz, R., Beyer, M. & Kuhn, K. A. (2007). Semantic integration in healthcare networks, *International Journal of Medical Informatics* 76(2-3): 201 – 207. Connecting Medical Informatics and Bio-Informatics - MIE 2005.

Liao, M., Hinkelmann, K., Abecker, A. & Sintek, M. (1999). A competence knowledge base system as part of the organizational memory, *XPS '99: Proceedings of the 5th Biannual German Conference on Knowledge-Based Systems*, Springer-Verlag, London, UK, pp. 125–137.

Manola, F. & Miller, E. (eds) (2004). *RDF Primer*, W3C Recommendation, World Wide Web Consortium.
URL: http://www.w3.org/TR/rdf-primer/

Passin, T. B. (2004). *Explorer's Guide to the Semantic Web*, Manning Publications Co.

Stader, J. & Macintosh, A. (1999). Capability modelling and knowledge management, *Applications and Innovations in Expert Systems VII, Proceedings of ES 99 the 19th International Conference of the BCS Specialist Group on Knowledge-Based Systems and Applied Artificial Intelligence*, Springer-Verlag, Berlin, pp. 33–50.

Sure, Y., Maedche, A. & Staab, S. (2000). Leveraging corporate skill knowledge - from proper to ontoproper, *in* D. Mahling & U. Reimer (eds), *Proceedings of the Third International Conference on Practical Aspects of Knowledge Management. Basel, Switzerland, October 30-31, 2000*. http://www.research.swisslife.ch/pakm2000/.

20

Magic Mathematics Based on New Matrix Transformations (2D and 3D) for Interdisciplinary Physics, Mathematics, Engineering and Energy Management

Prof. Dr.-Ing. Wolfram Stanek[1,2] and Dipl. Ing. Maralo Sinaga[3]
[1]*University of Applied Sciences Koblenz,*
[2]*Guest Lecturer at Swiss German University, BSD City-Jakarta*
[3]*Head of Mechatronics Department, Swiss German University, BSD City-Jakarta*
[1]*Germany*
[2,3]*Indonesia*

1. Introduction

Mathematics is magic. If we can either use one formula for a wide range of applications or the formula itself will produce magic properties. As one of several introductory examples the generally not well known Leibniz formula for calculating determinants in matrix theory will show that both the well known Laplace laws and Sarrus rules for evaluating matrices are only graphically visualised subsets of this ingenious Leibniz formula. Visualising complex formulas and matrix transformations in 2D and 3D as equivalent graphs is a basic method of the main author in this publication. The huge range of fascinating technical applications based on 2D magic matrices will be sketched: Constant distribution in all directions of numbers, power, energies, element properties, transport, automation, information flows etc or compensation of punctual disturbances without variation of sum of energy or automatic minimization of energy loss remaining constant distribution or both concentration of energies in near field and hiding of energies in far field or solving magic equation systems in mathematics without using back tracking methods etc.

2. Background

The extremely complex problem in mathematics of finding a perfect solution of a 4x4x4 - 3D - magic cube (64 unknowns, but 76 equations/conditions) with constant sum in all directions and continuous numbers from 1 to 64 was solved first by the German mathematician W. Trump in the year 2004 (Spectrum of Science, 2008-2): But this world wide first solution of a 4x4x4 magic cube was only based on parallel computations with several computers and extremely time-intensive back-tracking methods with time consuming solution. In contrast to this computer-based solution of 4x4x4 magic cubes in 2004, the main author Prof. Dr. W. Stanek has shown a new analytical method manually solving this problem during a presentation on German MemoMasters 2008 and 2009: Using this analytical method for 4x4x4

magic cubes, the manual 3D solution lasts a few minutes - applying this algorithm the solution time with MATLAB® needs only fractions of seconds (ca. 0.01 s).
The results of these matrix transformations for magic 64-cells-cubes show two main aspects:

a. Extremely fast solution of such matrix problems in 3D by immediate transformation from magic 2D matrices to magic 3D cubes with remaining central magic properties.
b. New idea solving large sets of linear equations (with also determinant-zero-matrix-property) NOT using conventional equation solvers (Gauss-Seidel, Newton-Raphson etc) and backtracking methods but only simplified geometrical 3D transformations and logic. This magic math algorithm is shown by visualised graph transformations and underlying equivalent structures.

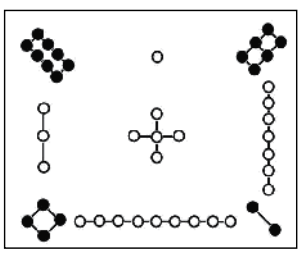
3. Magic square and magic cube (2800 B.C – 2008)

Magic squares were known to Chinese mathematicians, and Arab mathematicians, possibly as early as the 7th century, when the Arabs conquered northwestern parts of the Indian subcontinent and to learned Indian mathematicians and astronomers, including other aspects of combinatorial mathematics. The most famous 2D magic squares are the Lo Shu square and the Duerer square.
A normal magic square contains the integers from 1 to n^2. The constant sum in every row, column and diagonal is called the magic constant or magic sum, S. The magic constant of a normal magic square with continuous numbers depends only on n and has the value:

$$S = \frac{n}{2}\left(n^2 + 1\right) \tag{1}$$

3.1 Lo Shu Square

Lo Shu square or the **Nine Halls Diagram** is the unique normal magic square of order 3 x 3. Lo Shu is part of the legacy of the most ancient Chinese mathematical and divinator traditions, and is an important emblem in "Feng Shui" the art of geomancy concerned with the placement of objects in relation to the flow of 'natural energy'.
The Lho Shu square was introduced in 2800 BC. Fig. 1.(a) shows the Loh Shu square used symbolism instead of numbers, and Fig. 1.(b) representing continuous number 1 to 9 this square. The Loh Shu square dimension is n=3, then the magic sum S is 15.



(a) Loh Shu square with symbols

8	1	6
3	5	7
4	9	2

(b) Loh Shu square with numbers

Fig. 1. Loh Shu square (Wikipedia, 2010)

3.2 Duerer magic matrix

The Renaissance engraving "Melancholia I" was developed by the German artist, painter, and mathematician Albrecht Duerer (in the year 1514). This image is filled with mathematical symbolism and in the upper right corner of the first picture a square can be seen. The Fig. 2(a) shows an enlarged view of the Duerer's square cropped from the image. This square is known as a magic square and was believed by many in Duerer's time to have genuinely magical properties. It does turn out to have some fascinating characteristics worth exploring.

(a)

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

(b)

Fig. 2. Duerer square, Melancholia I, 1514, (Wikipedia, 2010)

The Duerer's square in Fig. 2(b) is filled up with continuous numbers 1 to 16. The square dimension is n=4, then the magic sum S of Duerer's square is 34.

3.3 Sudoku

Sudoku is a logic-based, combinatorial number-placement puzzle. The objective is to fill a 9×9 grid with digits so that each column, each row, and each of the nine 3×3 sub-grids that compose the grid contain all of the digits from 1 to 9. The puzzle setter provides a partially completed grid, which typically has a unique solution.

Completed puzzles are always a type of Latin square with an additional constraint on the contents of individual regions. For example, the same single integer may not appear twice

- in the same 9x9 playing board row
- in the same 9x9 playing board column or
- in any of the nine 3x3 subregions of the 9x9 playing board.

The puzzle was popularized in 1986 by a Japanese puzzle company and became an international hit.

An extended example of the Sudoku game is shown in the Fig. 3, where a 9x9 square **A** is developed from a 3x3 square. Then the other cells are filled by using a regular shifting

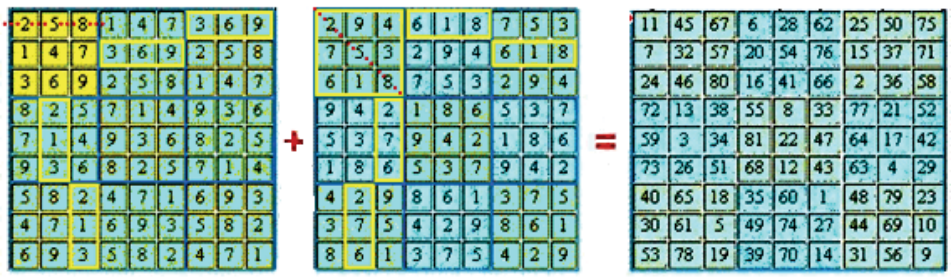


Fig. 3. Sudoku 9x9 from 3x3 square and bi-magic square (www.multimagie.com, 2009)

method. From the square **A**, then square **B** is constructed by using rotating technique, and finally the bi-magic square **C** (with continuous number 1 to 81) is developed through addition of **A** and **B,** example applying the formula **C=9 · (A–1)+B**. This magic square is bi-magic (or multi-magic) if it remains magic after each of its numbers have been squared. This was introduced by Tarry and Cazalas. All cells content are squared, resulting in magic sums of C=369 and D=20049.

4. Components of creative intelligence, Leibniz matrix and new solution technique.

The following MATLAB® program cutoff to calculate the determinant of n x n matrix should be a central background of this publication. According to the famous Leibniz formula for determinant calculation of any n x n matrix, the fact shows that the Method of Sarrus and Method of Laplace to solve the determinant of the matrix use similar concept of the Leibniz formula with different visualisations. Both Sarrus and Laplace Methods could be structurised and visualised in mnemotechnique method by the main author. Finding the determinant of nxn matrix using the Leibniz formula is shown in the equation (2):

$$\det(A) = \sum \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} \right) \tag{2}$$

Based on this point, W.Stanek provides an algorithm to obtain any 3D magic structure, especially 64 cells cube, primarily the visualised structure of the algorithm. At the same time it will be shown that in a linear equation with a lot of unknowns (64 unknowns and 76 equations) is first solved by using the logical method and visual solution. This problem was first solved in 2004 by using computers, working in parallel and based on the back tracking algorithm method.

The equation is represented in MATLAB® function and algorithm shown as:

```
% MATLAB and Leibniz formula provide the same result
% det(M1) = -360  for 3x3 Loh Shu Matrix
 % For magic 4x4 Duerer Matrix both methods yield det(M2) = 0, too

detA_Matlab = det(M1)

% Leibniz-Formula for all nxn-Matrices, here only shown for 4x4-Matrix :
% With n=4 follow  4! = 1x2x3x4 = 24 Terms for solutions of det(A)
% With a14=0; a24=0; a34=0; a44=1; a41=0; a42=0; a43=0;
% Leibniz formula also for 3x3-Matrices like i.e. magic 3x3 Loh Shu Matrix.

detA_Leibniz = (a11*a22*a33*a44 + a11*a23*a34*a42 + a11*a24*a32*a43...
- a11*a22*a34*a43 - a11*a23*a32*a44 - a11*a24*a33*a42...
+ a12*a21*a34*a43 + a12*a23*a31*a44 + a12*a24*a33*a41...
- a12*a21*a33*a44 - a12*a23*a34*a41 - a12*a24*a31*a43...
+ a13*a21*a32*a44 + a13*a22*a34*a41 + a13*a24*a31*a42...
- a13*a21*a34*a42 - a13*a22*a31*a44 - a13*a24*a32*a41...
+ a14*a21*a33*a42 + a14*a22*a31*a43 + a14*a23*a32*a41...
- a14*a21*a32*a43 - a14*a22*a33*a41 - a14*a23*a31*a42)

% NOTE: Leibniz is central for all det(A)-calculations by Sarrus and by Laplace
```

```
% 3x3 matrices calculated by Sarrus Rule directly from Leibniz Formula
% nxn matrices (n=3, 4, ...) by Laplace Rule directly from Leibniz, too
% Both rules of Sarrus and Laplace are visualised structures of  the
% Leibniz Formula detA_Leibniz (above shown for 4x4-Matrices)
% This Leibniz Formula is ingenious as basis for Sarrus, Laplace etc
```

From equation (2) following equation (3) can be derived.
It is possible to expand a determinant along a row or column using this formula, which is efficient for relatively small matrices. To do this along row *i*, say, we write:

$$\det(A) = \sum_{j=1}^{n} A_{i,j}\, C_{i,j} = \sum_{j=1}^{n} A_{i,j} \cdot (-1)^{i+j} \cdot M_{i,j} \tag{3}$$

where the $\boldsymbol{C}_{i,j}$ represents the *i,j* element of the matrix cofactors, i.e. $\boldsymbol{C}_{i,j}$ is $(-1)^{i+j}$ times the minor $\boldsymbol{M}_{i,j}$, which is the determinant of the matrix that results from $\boldsymbol{A}$ by removing the *i*-th row and the *j*-th column, and *n* is the length of the matrix.
The determinant of a 2x2 matrix A is calculated by:

$$M = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \det(A) = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}$$

$$\det(A) = \left(a_{11} \cdot a_{22} - a_{21} \cdot a_{12}\right)$$

For a of 3x3 matrix, the determinant is calculated by using the Sarrus method, derivated from Leibniz's formula:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad \det(A) = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Vmatrix}$$

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{vmatrix} \quad - \quad - \quad - \quad + \quad + \quad +$$

$$\det(A) = \left(a_{11} \cdot a_{22} \cdot a_{33} + a_{12} \cdot a_{23} \cdot a_{31} + a_{13} \cdot a_{21} \cdot a_{32} - a_{31} \cdot a_{22} \cdot a_{13} - a_{32} \cdot a_{23} \cdot a_{11} - a_{33} \cdot a_{21} \cdot a_{12}\right)$$

To find the determinant of a 3x3 matrix according the Laplace formula shown in the equation (3):

$$\det(A) = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Vmatrix} = a_{11} \begin{Vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{Vmatrix} - a_{12} \begin{Vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{Vmatrix} + a_{13} \begin{Vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{Vmatrix}$$

$$\det(A) = \left\{ a_{11}\left(a_{22}\cdot a_{33} - a_{32}\cdot a_{23}\right) - a_{12}\left(a_{21}\cdot a_{33} - a_{31}\cdot a_{33}\right) + a_{13}\left(a_{21}\cdot a_{32} - a_{31}\cdot a_{22}\right)\right\} \quad (4)$$

Because it is dealing with a 3× 3 matrix, it sets up the 3× 3 sign matrix. This is always a "checkerboard" matrix that begins with a "+" sign in the upper left corner and then alternates signs along rows and columns.

The Leibniz formula is the root of the Sarrus formula and the Laplace formula. The regularity of the Leibniz-, Laplace-, and Sarrus-determinant calculation was the basis for the main author developing magic matrices and cubes through visualised transformation of the cell contents (shifting, rotating and reflecting or mirroring).

5. Computer solution

It was only possible to solve a 4x4x4 cube (76 equations with 64 unknown) by using a several computers, working in parallel, and based on the backtracking algorithm method. This was shown by the German mathematician Walter Trump in year 2004.

New Idea

The 4x4x4 cube (76 equations with 64 unknown) will be solved now by using only logical thinking and geometrical methods (bending surfaces). Assume a box, with 4x4 cells on each side is opened into a 2 dimensional plane as shown in the Figure 4; the number must be match each other to the side plane. The following sequence is used to solve the magic-matrix and cube respectively.

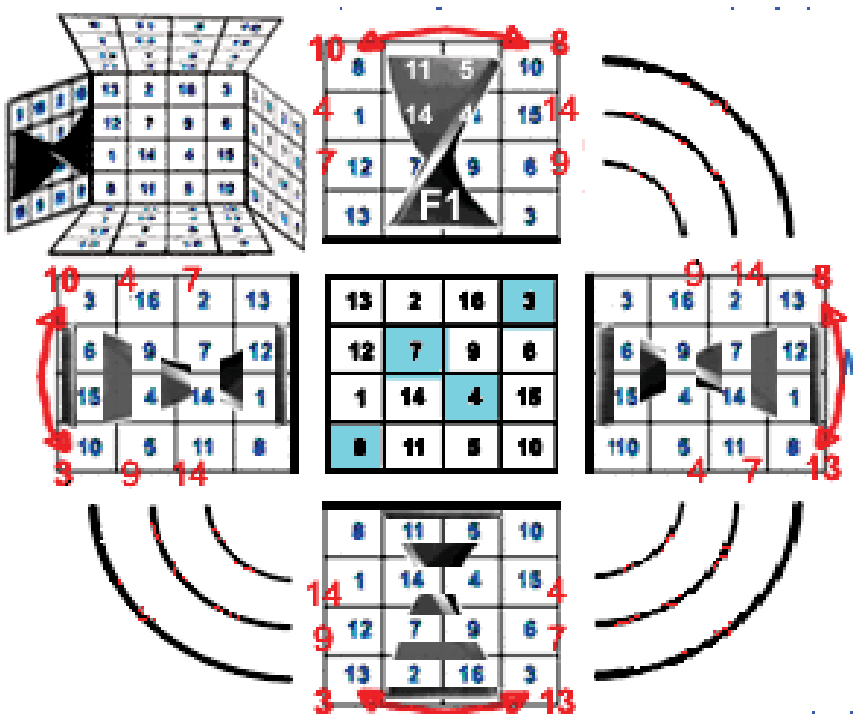


Fig. 4. Solution idea based on "box-exploding" (Stanek, 2009)

Step 1. Start with any magic 4x4 matrix M1,
Step 2. M1 is reflected in all sides of the box.
Step 3. Use the logic method, match all the edge cells
Step 4. From M1 until M4, magic cube 2 is constructed using surface transformation, bending, or reflecting (mirroring).

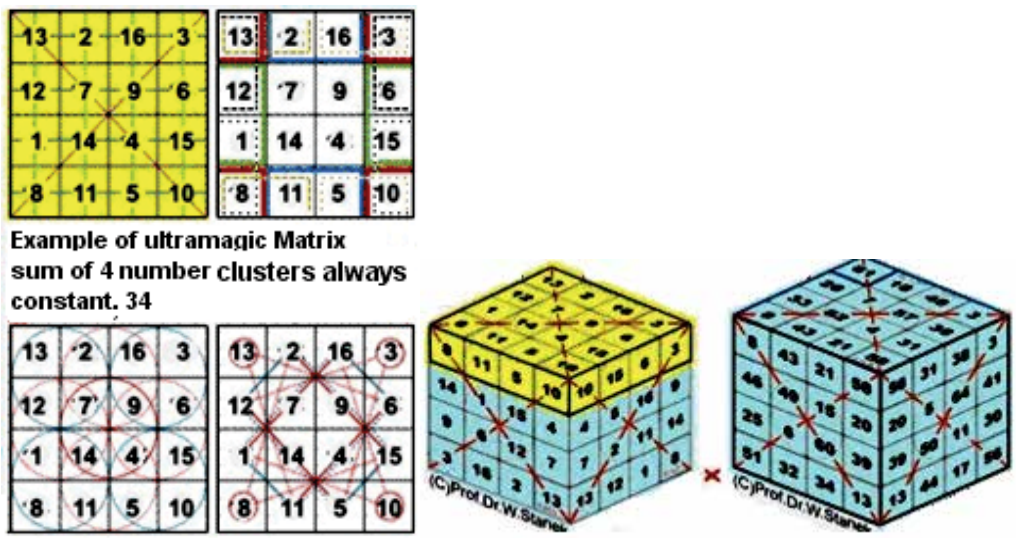
The computer based magic cube solution, which is discovered in 2004 with highest degree of perfection is first in 2008 analytically solved by the main author. The solution ideas and the important steps using the Stanek Method to solve the magic cube will be shown in the following pages and a pattern solution is attached.

6. Short information: magic cube with Stanek-Method analysis

In the above shown graphical method the solution of magic cube is explained.
All related data of „Sudoku to the power of 3"application will be simple and always enough to solve or to construct in the 4 main layers.

6.1 Example: magic + ultra-magic square and cube

A magic square is bi-magic (or multi-magic) if it remains magic after each of its number has been squared and an ultra magic square has more extended properties. The following 4x4 square shows an example of ultra magic square (Fig. 5.(a)).



(a) (b) (c)

Fig. 5. (a) Perfect square or ultra-magic square with continuous number 1 to 16 all rows, columns, diagonals, and four neighbor cells always resulting magic sums = 34
(b) Magic Cube developed from magic square, with continuous number 1 to 16 in 3D, resulting a perfect magic sum=34 in rows, columns, diagonals
(c) Magic Cube developed from magic square, with number 1 to 64 in 3D, resulting a perfect magic sum=130 in rows, columns, diagonals in x-,y-, and z-planes (Stanek, 2009)

The sum of all numbers in the horizontal, the vertical as well as in the diagonals are equal and also the sum of all four neighbor cells, which form a square are always constant to 34. The square pattern in Duerer's square and Loh Shu's square are easy to memorised. Starting with predetermined Loh Shu's pattern, a 9x9 squares can be constructed easily by applying a shifting, reflecting method.
A magic cube with 4x4x4 dimension can be developed for instance from the known Duerer's square pattern, or extended from ultra magic square, converted by reflecting, shifting and bending.

6.2 An ultramagic square as the base of Stanek Cube Developments

Each of 48 given ultra-magic matrices or squares results in a new magic cube with a maximum possible degree of perfection. Through simple transformation, (shifting, rotating and reflecting) it is easy to construct other ultra magic matrices.

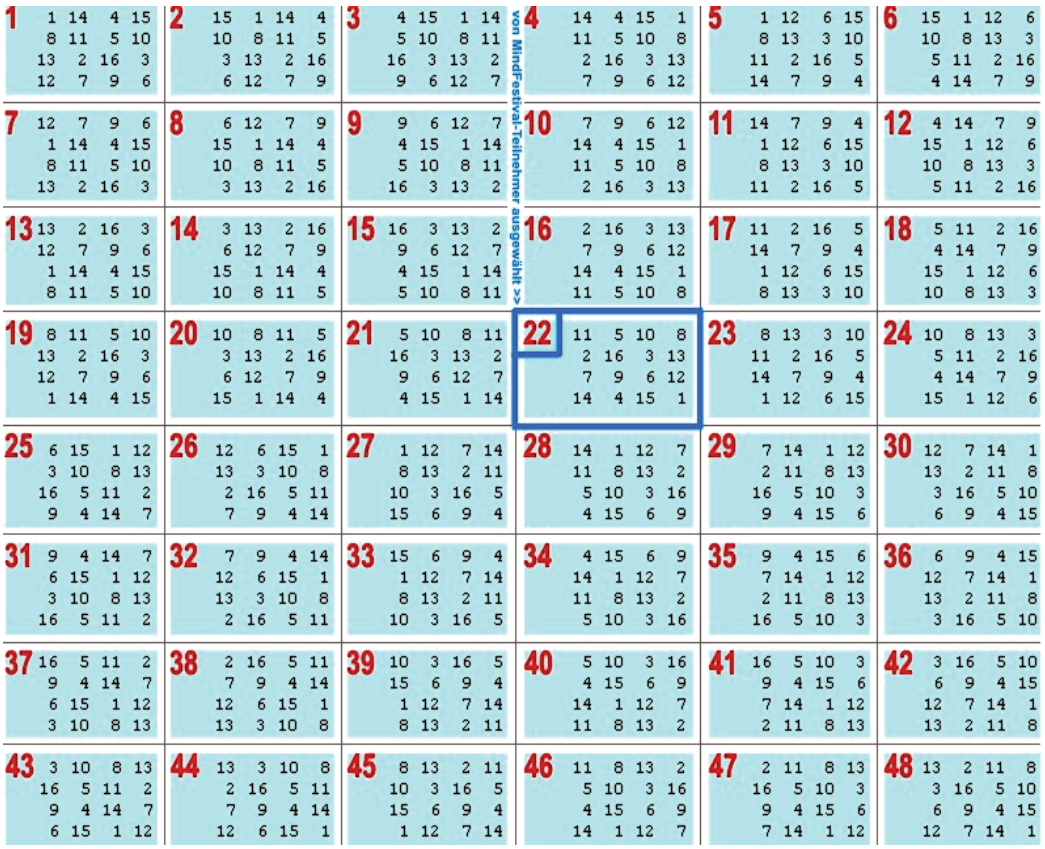


Fig. 5. 48 possible magic squares, constructed through shifting, rotating and reflecting.

7. Solution by using mnemonic scheme for Stanek Cube No.1 + No.2

On the Fig. 4 it is shown how to generate and to develop a magic square and magic cubes from a given ultra-magic square. An example is the square number 22 shown in the Fig. 5.

(given by the audience at MemoMasters – MindFestival 2009 to Prof. W. Stanek for manual solution). Starting by choosing this ultra magic square number 22 as the start matrix M1, the next ultra magic square M2 is created by using the reflecting (mirroring) method. Then from matrix M2 we can construct the next matrix M3 by transposing and at the same time reflecting the content of cells and finally the matrix M4 is generated from M3 using reflecting in each four neighbor cells (number 1 to 16).
The resulting ultra-magic matrices M1 until M4 are used to develop the next layers; creating the magic cube layers with a sketching pattern (+0, +16, +32, +48), as shown in the Fig.6.

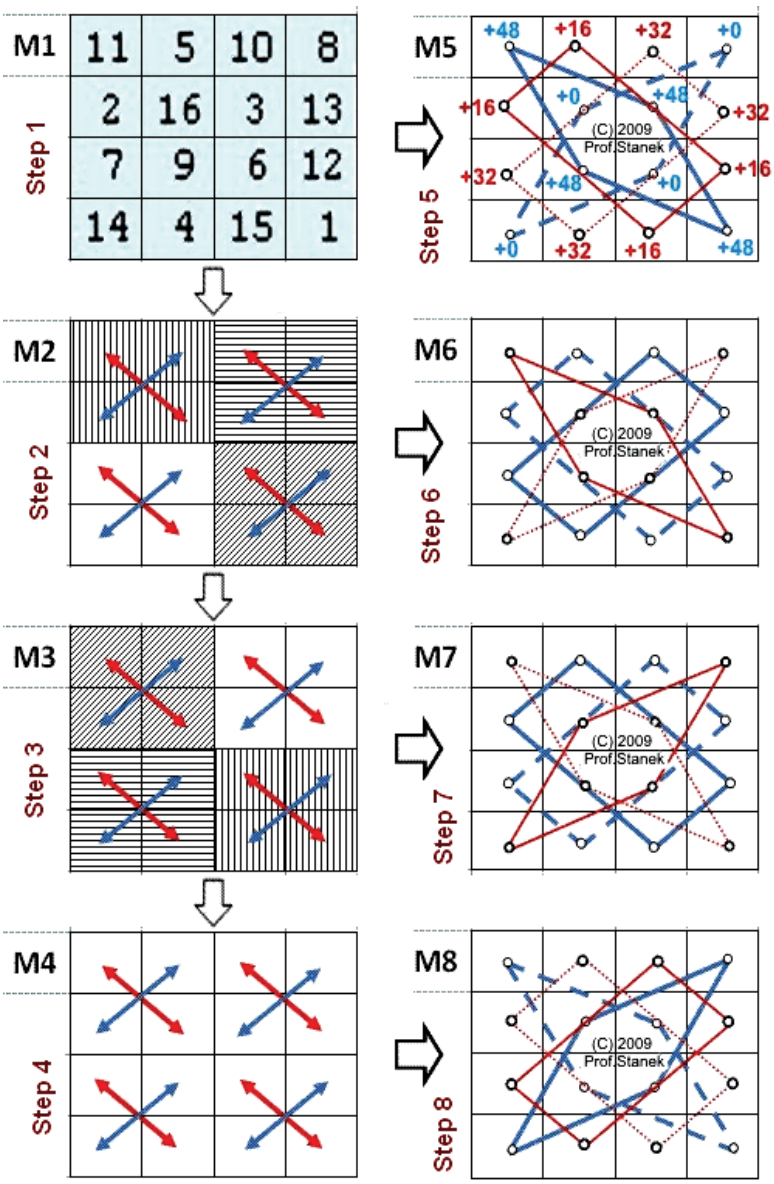


Fig. 6. Pattern Solution to solve a magic cube (Stanek, 2009)

7.1 First comparison: computer and analytical solution with highest degree of perfection

A comparison between a solution which was reached by using computer backtracking method (W. Trump, 2004) and an analytical solution using logic and brain memory shows that the Stanek analytical solution delivers the highest perfect precision of magic cube (MemoMasters 2008).

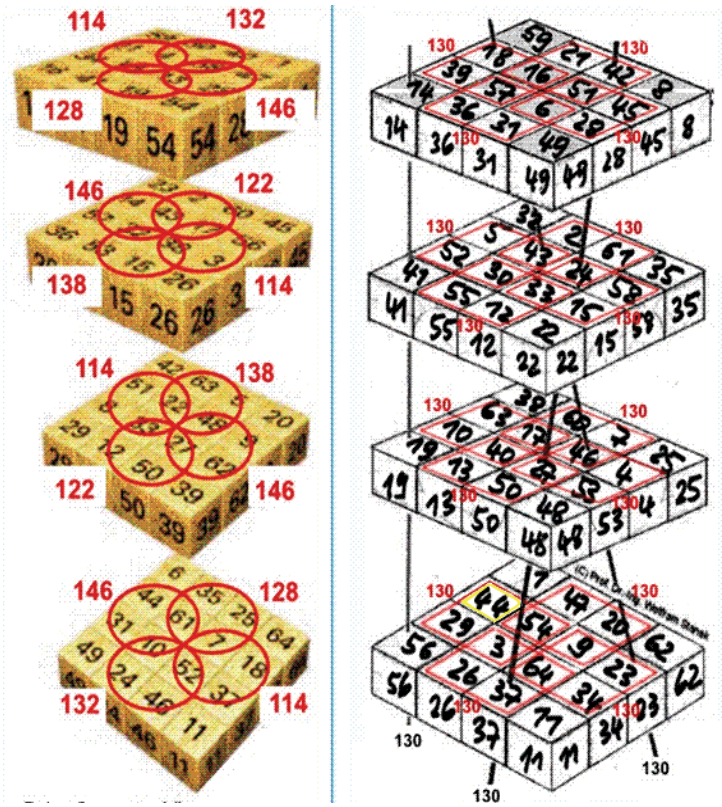


Fig. 7. (a) Magic cube, computer solution by W.Trump (Spektrum Wissenschaft, 2008)
(b) Analytical solution by Prof. Wolfram Stanek.

This result for the specific analytical cube was predictable, since a start ultra-magic matrix was chosen. The result in Fig. 7.(a) was shown in "Spektrum der Wissenschaft", 2008 and the Fig. 7.(b) was represented on MindFestival in August 2009. In all the transformations of planes in the space partially some magical sum properties are lost. If many magical properties in the initial matrix are available (at ultra magic matrix of the case), the result of the magic cube (1-64) has a relatively highest degree of perfection.

7.2 Second comparison: minimising the deviation of space diagonals with computer and analytical methods

By choosing the Duerer's matrix, which is not perfect, a magic matrix is created as the start matrix in developing other magic cubes as shown in the Fig. 8.

The Fig.9 shows the difference between W.Trump's magic cube which is solved by computer and the main author's analytical solution by starting with the magic matrix in the

previous Fig. 8. The gray areas in the figure represent the difference between both solutions. In the right column is shown the magic cube, which is constructed through twisted matrix transformation. The highlighted cells (rows, column and diagonals) produce the constant sum of the magic cube.

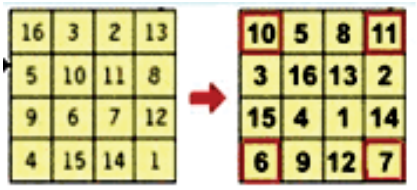


Fig. 8. Constructing a magic cube, starting with Duerer's matrix (Stanek, 2009)

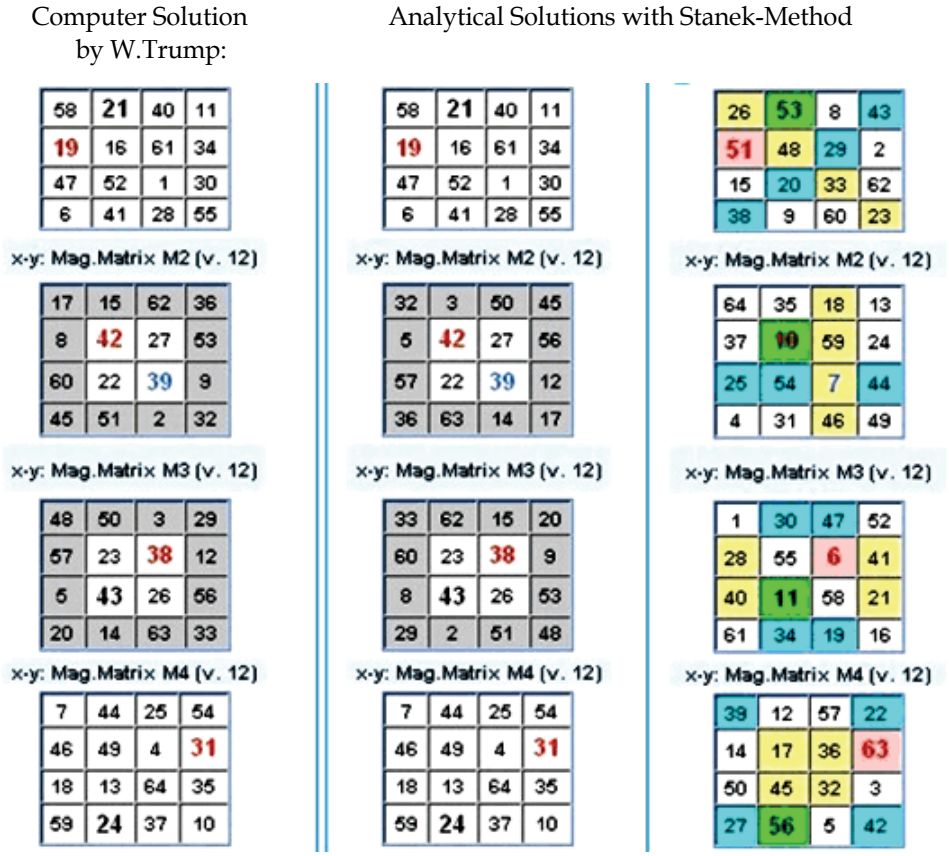
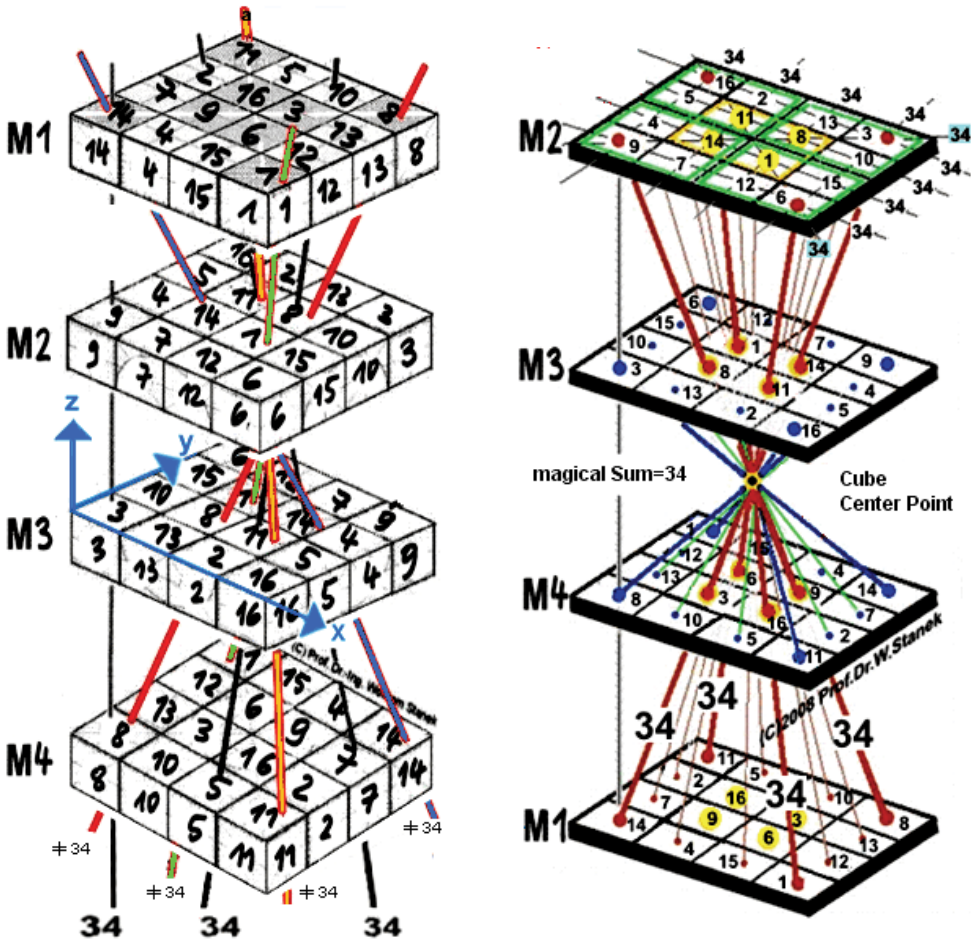


Fig. 9. Complete solution to build magic matrices and magic cube with highest symmetry (Stanek, 2009)

The comparison also shows that the Stanek analytical method works magically not only in start-matrix of ultra-magic, but also in normal magical start squares, which have at least partly-some pandiagonal magic properties. In Duerer's matrix the major side-diagonals (e.g. 2 +10 +9 +1 = 22, etc) are not magical, but parts of the side-diagonal always have a magic constant (e.g. 3 +5 +14 +12 = 34, etc).

8. New magic matrix applications for interdisciplinary system design

Now, the question is: "Is this a mathematical phenomenon, or what we see and what we need?" A main question for our brain and our life is what is important and not important for us temporarily? According to "Spektrum der Wissenschaft, 2008-2", there is no existing absolutely perfect cube 4x4x4 because the main spatial diagonal always deviates.
Is it correct or is it a starting question? A near perfect 4x4x4 cube is shown in the Fig.10 (Nintendo MemoMasters 2009 –MindFestival)



(a) (b)

Fig. 10. (a). Near perfect magic cube (b) Perfect magic cube (Stanek, 2009)

The only once calculated building block - depending on the "Position" -surprisingly contains quite different symmetry results.

- In the Fig. 10.(a): in all mutually perpendicular surfaces of magic numbers (symmetry) there is 'perfect Cartesian', but not in the spatial diagonals.
- In the Fig. 10.(b): just by shifting the layer M1 "perfect central symmetry" is achieved. All major and minor diagonals have a constant sum of 34 (suggestion by Einstein "... all is relative")

8.1 Strategy-checks "Sudoku to the Power of X" + scientific and technological point of view.

(A) "Sudoku to the power of 3"with ultra-magic Start-squares & Stanek-Method.

1. Constructing 11 new magic squares from 1 ultra magic square; select any ultra magic square from Fig. 5 on page 7 (manually, required time: less than 90 seconds)
2. Constructing near perfect magic cube no. 1 (number 1 to 16) from point A.1.
3. Constructing any possible perfect magic cube no.2 (number 1 to 64) from point A.2
   Notes:
   Until 2004 this mathematical problem was thought to be unsolvable, then in 2004 it was solved by Trump using parallel computing of computers and backtracking method; in 2008 this problem was solved analytically and finally in 2009 it is solved by using the Stanek-Pattern Strategy during MemoMasters 2009.

For Strategy B and Strategy C improvement, try your own "creative intelligence " in logic and mathematic areas.

Try to select a solution strategy from the following improvement that you can find quickly.

(B) "Sudoku to the power of 3": Additional Strategy to 4x4x4-Stanek-Cube and 3x3x3-Rubik-cube

1. Try to find an extension strategy for a Stanek magic cube with a given sum value greater than 130; example the magic sum equals 274 and the maximal number in the cube equals 100.
2. Construct with this extension (just addition) the new cube directly from point A.3
3. Try to apply the Stanek-strategy to generate an arbitrary sum-number (see page 5) for the color rubik cube (divisible by 3).

(C) "Sudoku to the power of 2" = Combination of Sudoku and interesting special magic squares

1. Constructing a magic 4x4-Sudoku X with four 2x2-fields (with each number 1- 4)
2. Apply the 3x3-Loh-Shu-square, included design rule, so that each odd-square nxn (n = 3, 5, 7, 9, ... etc) can be constructed immediately.
   Note: Position of Serial numbers 1-9 including the solution scheme. Start-number 1 is always above in the middle, No. 2 in the right column below.
3. Construct a 9x9 magic square Y with this method from C.2: All rows, columns and main diagonals have a constant sum

$$S = n \times (n \times n + 1) / 2 \text{ for number from 1 to } n^2.$$

   Example, n=9, then S=396, etc.
   For magic cube magic sum

$$S = n \times (n \times n \times n + 1) / 2, \text{ for number from 1 to } n^3 \quad (5)$$

   example 4x4x4- cube, n=4 then S = 130, etc
4. Try to prepare a 4x4 magic square (magic or ultra-magic square) so that for any number, the sum of all numbers in all rows, columns and main diagonals are equal.

5.a). Construct a 9x9-magic Sudoku A from a 3x3-Loh-Shu-square.

5.b).Construct a 9x9-magic Sudoku B from Sudoku A in 5.a).

5.c). Construct bi-magic square C (Number 1 to 81) from Sudoku A and Sudoku B in 5.a) and 5.b).

5.d). Construct a new magic square D, where all the 81 number of square C in 5.c) are squared.

NOTE:

a. Magic "Sudoku to the power of 3" cube number 1 and 2 is relatively easy to solve by using Stanek's predetermined pattern Strategy-Method; otherwise, without this strategy the cube can be only solved by using a computer program.
b. With the Stanek-Method, a linear equation system (76 equations with 64 unknown) is solved primarily with logic and geometry.
c. The mathematical genius, physicist and astronomer, Galileo Galilei (1564-1642), said, "Who understands geometry can understand the world"
d. Problem with difficult creativity are not possible to be solved without strategy, even with best memory.
e. Test your own understanding strategy using the "Sudoku to the power of X"- Checks **(B)** & **(C),** and your temporary capability for Logic-Mathematic Strategy in interdisciplinary area "Creative Intelligence".

Application range of "Magic Cube" in Science and Technology.

This is a short overview of mostly unused fantastic magic application with high precision in matrix-areas in 2D and 3D in Science and Technology Point of View.
Aspects of R&D in engineering with focus on mechatronics and integrated interdisciplinary differential equations are extended matrix operations with magic matrix techniques too.
Unknown applications of magic matrix could be:

1. Optimised magic distribution of energy, power, element properties, information fluxes, etc. constant in all directions in 2D or 3D.
2. Automatic Minimising of energy losses in all directions in 2D or 3D.
3. Direct compensation of punctual disturbances with unchanged sum of energy
4. Optimised logistic automation and transport with different motors.
5. Magic matrices as explanation of chemical structures or unsolved problems in physics.
6. A quick and new way in solving undetermined systems of equations without using conventional method (iterative solvers and backtracking method).

8.2 Tests of Stanek Algorithm for magic 4x4x4-cubes

To understand the analytical algorithm of magic 4x4x4 cube, the following sequence is used to develop a magic cube;

Step 1. Create matrix M1, from the 4x4 magic matrix i.e. numbers 22 given in the Fig. 5.

$$M1 = \begin{bmatrix} 11 & 5 & 10 & 8 \\ 2 & 16 & 3 & 13 \\ 7 & 9 & 6 & 12 \\ 14 & 4 & 15 & 1 \end{bmatrix} \qquad M5 = \begin{bmatrix} 59 & 21 & 42 & 8 \\ 18 & 16 & 51 & 45 \\ 39 & 57 & 6 & 28 \\ 14 & 36 & 31 & 49 \end{bmatrix}$$

Step 2. Construct the matrix M2 from M1 using reflection of the cells content diagonally (the four neighbor cells of magic matrix) according to the pattern solution shown in the Fig. 6.

$$M2 = \begin{bmatrix} 16 & 2 & 13 & 3 \\ 5 & 11 & 8 & 10 \\ 4 & 14 & 1 & 15 \\ 9 & 7 & 12 & 6 \end{bmatrix} \quad M6 = \begin{bmatrix} 32 & 2 & 61 & 35 \\ 5 & 43 & 24 & 58 \\ 52 & 30 & 33 & 15 \\ 41 & 55 & 12 & 22 \end{bmatrix}$$

Step 3. Developing the matrix M3 from M2 by reflection of the internal four neighbor cells diagonally.

$$M3 = \begin{bmatrix} 6 & 12 & 7 & 9 \\ 15 & 1 & 14 & 4 \\ 10 & 8 & 11 & 5 \\ 3 & 13 & 2 & 16 \end{bmatrix} \quad M7 = \begin{bmatrix} 38 & 60 & 7 & 25 \\ 63 & 17 & 46 & 4 \\ 10 & 40 & 27 & 53 \\ 19 & 13 & 50 & 48 \end{bmatrix}$$

Step 4. Developing the matrix M4 from M3 by using reflection of the cells content diagonally (the four neighbor cells of magic matrix), according to the pattern solution shown in the Fig. 6 and similar to the step 2.

$$M4 = \begin{bmatrix} 1 & 15 & 4 & 14 \\ 12 & 6 & 9 & 7 \\ 13 & 3 & 16 & 2 \\ 8 & 10 & 5 & 11 \end{bmatrix} \quad M8 = \begin{bmatrix} 1 & 47 & 20 & 62 \\ 44 & 54 & 9 & 23 \\ 29 & 3 & 64 & 34 \\ 56 & 26 & 37 & 11 \end{bmatrix}$$

NOTES:

- The matrices M1 until M4 are magic-matrices with the cells content of number 1 to 16.
- To construct 3D magic cube, layer 1 (M5), layer 2(M6), layer 3 (M7) and layer 4 (M8) are developed from M1, M2, M3 and M4 respectively by using mnemotechnique scheme solution as shown in the Fig. 6.

The Fig. 11 shows a sample of MATLAB® program listing as a test algorithm of a magic 4x4x4 cube which is tested by Martin Eka Putra and Jemmy Dianto (SGU students) during internship in Germany, and F. Halfmann (Assistant in Lab Automation & Robotics FH Koblenz) and then the magic sum is represented graphically in the next Fig. 12. It can be clearly seen that the continuous numbers from 1 to 64 fit in the cube are without any repetition. Fig. 13 shows the sum of the columns, rows, and main diagonals in 3D equal to 130 and the spatial diagonals' sum also has symmetry.

Application of the magic matrix transformation in Science and Technology

Now, we can apply the concept of the magic matrix in 2D as well as in 3D into the real science and technology field. Let a production system consisting of four rows of transport mechanism (e.g. using conveyors) be assumed as a matrix. Each of conveyors are divided into four parts and driven by four electric motors, as illustrated in the Fig. 14 (a). How is the energy or power in all conveyor motors with different power distributed so that all four clusters can in parallel produce products with the same power respectively energy?

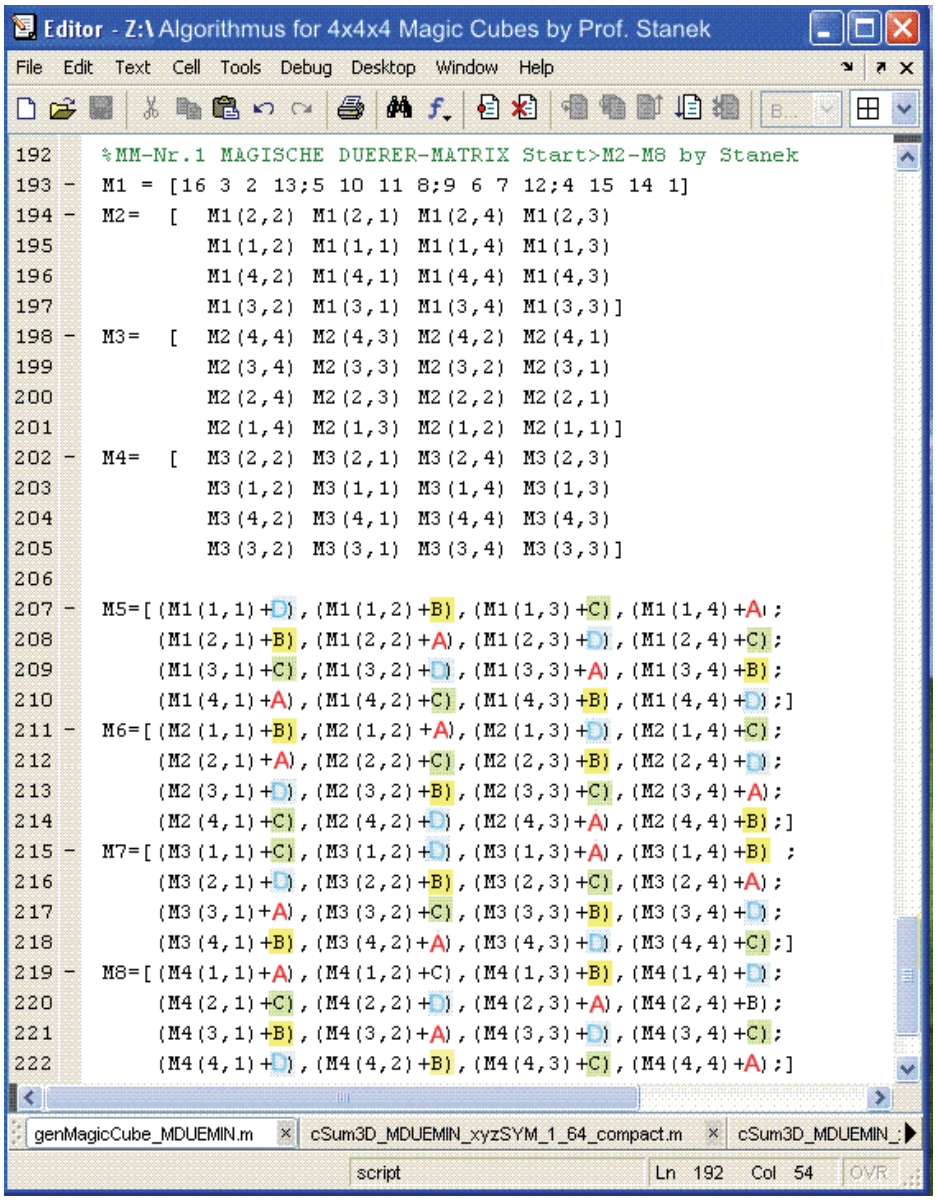


Fig. 11. MATLAB® program listing for test algorithm, with A, B, C, D equal to +0, +16, +32, +48. (Stanek, 2010)

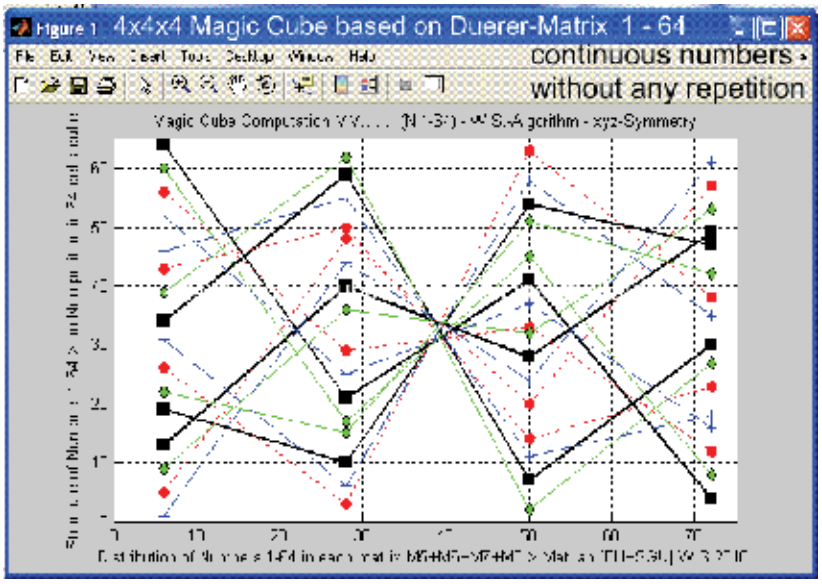


Fig. 12. Distribution of continuous number 1 to 64 in the magic cube (Stanek, 2010)

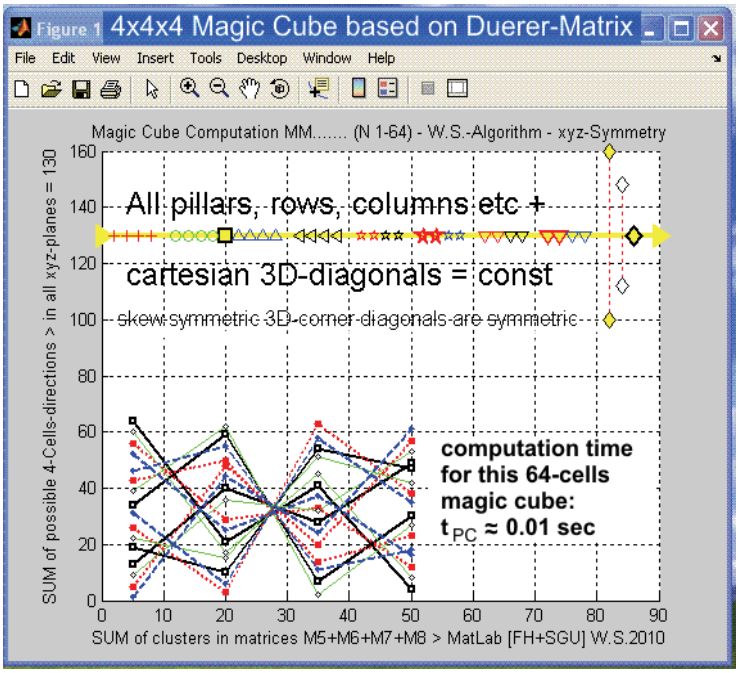
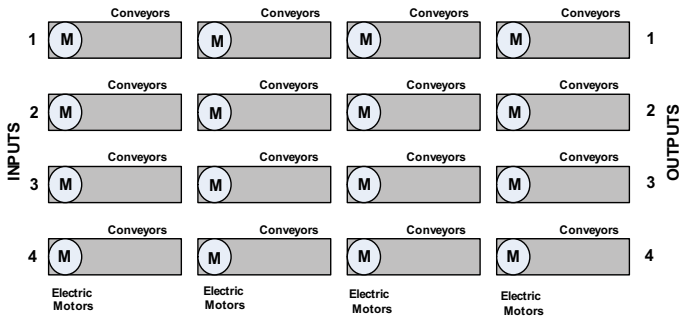


Fig. 13. Magic sum of rows, column, and diagonals (Stanek, 2010)



(a) Technology

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

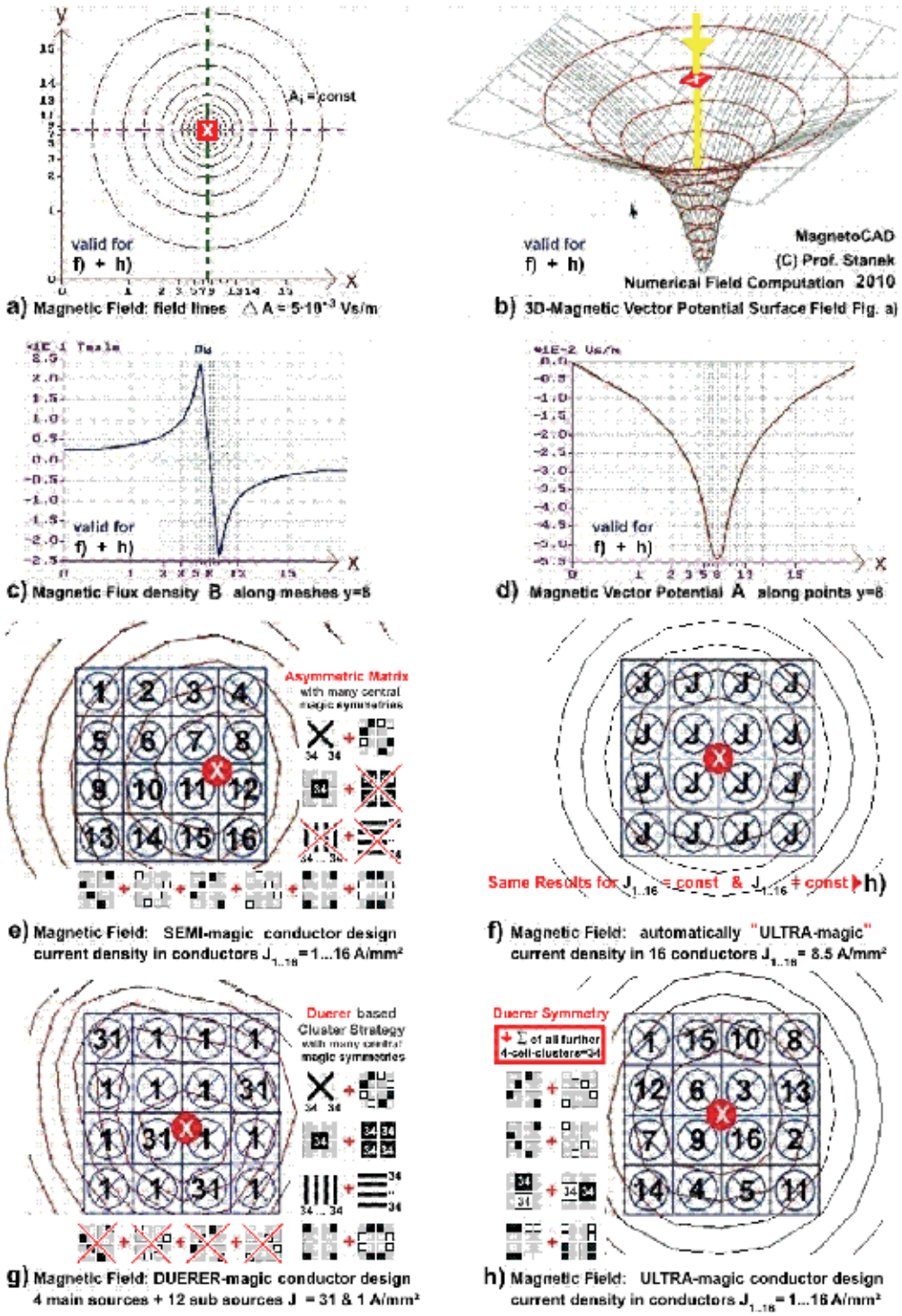
(b) Continuous power

104	6	81	9
8	82	7	103
5	101	10	84
83	11	102	4

(c) Clustered power

Fig. 14. (a) Example of magic square application in the production line, with the maximum power distribution per row and column less than 200kW (Stanek, 2010)
(b) Start Matrix, i.e. Duerer matrix, (c) Constructed magic matrix, shows the distributed power in each row using i.e. 4 main power drives and smaller sub-drives, column and diagonal are constant sum of 200kW.

To produce a constant amount of production in any row, how is the energy and motor power respectively distributed without exceeding maximum power of i.e. 200 kW, total maximum power consumption per conveyor-line is derived from Duerer start-matrix in Fig. 14 (b) and developed into power distribution as shown in the Fig. 14 (c).
An example of a 4x4 magic square with constant sum zero, known as zero-sum matrix, is shows in the Fig. 16.
The represented matrix is derived from a magic matrix by addition and subtracting of cell contents with a constant, so the resulting matrix has a magic constant sum zero. This can be applied to the design of coils or transformer windings as shown in the next Fig. 15.Another example is if we apply the magic square 2D to construct real 3D applications in magnetic field as shown in the Fig 15 and Fig. 16. It shows the symmetry with focus on separate x-, y-, and z-regions.
Real applications with totally different features in all cells of 3D can be optimised with constant sums in arbitrary directions using the real magic-matrix algorithm for 4x4x4 cubes shown in Fig. 6 to Fig. 10.
From the Fig. 15 (a) can be seen that the spatial field is zero due to the symmetry of the magic matrix in developing of the windings currents. The sum of all four-cell-clusters is always zero. This phenomenon might be theoretically interpreted as a galaxy black hole, too.
The magic square and cube could be applied also in other science and technology fields, such as energy management, flow management, logistics, and thermodynamics etc. See the spectrum of possible magic –matrix applications in section 7.1 too.



a) Magnetic Field: field lines $\Delta A = 5 \cdot 10^{-3}$ Vs/m
b) 3D-Magnetic Vector Potential Surface Field Fig. a)
c) Magnetic Flux density B along meshes y=8
d) Magnetic Vector Potential A along points y=8
e) Magnetic Field: SEMI-magic conductor design
current density in conductors $J_{1..16}$ = 1...16 A/mm²
f) Magnetic Field: automatically "ULTRA-magic"
current density in 16 conductors $J_{1..16}$ = 8.5 A/mm²
g) Magnetic Field: DUERER-magic conductor design
4 main sources + 12 sub sources J = 31 & 1 A/mm²
h) Magnetic Field: ULTRA-magic conductor design
current density in conductors $J_{1..16}$ = 1...16 A/mm²

Fig. 15. Current windings design with semi-magic & ultra-magic matrices

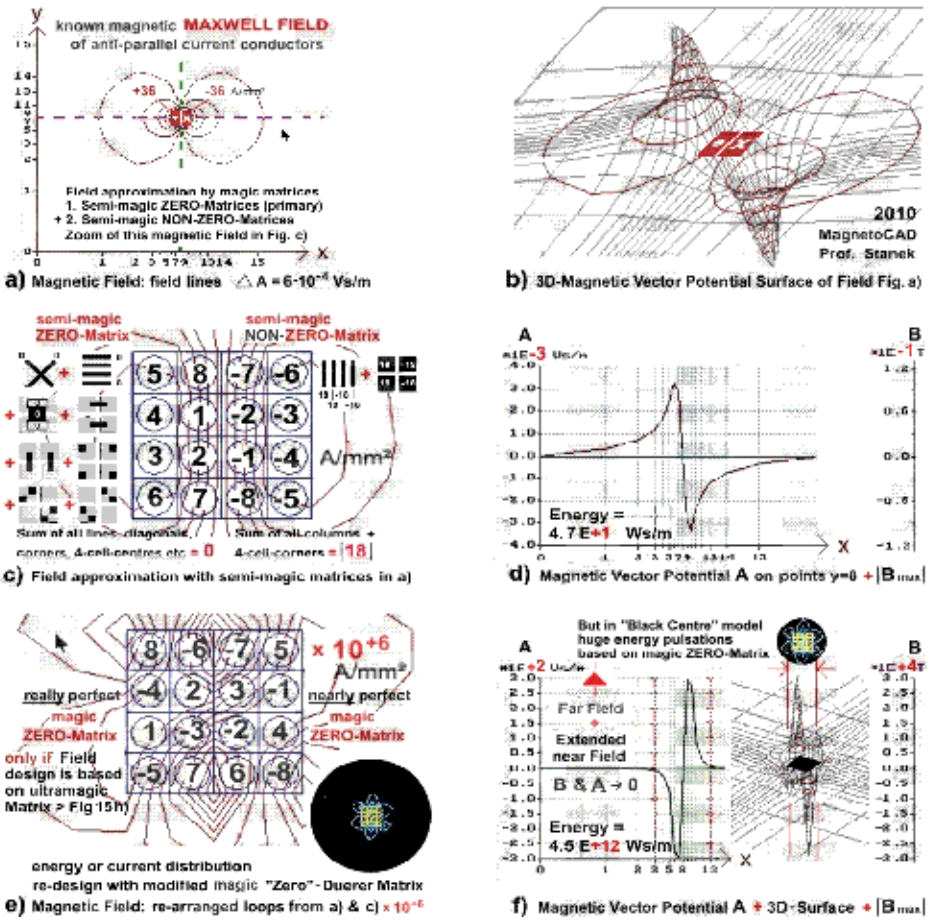


Fig. 16. Current windings and field design with magic ZERO matrices

9. Acknowledgements

Thanks of main author to his appreciated colleague and co-author Dipl. Ing. Maralo Sinaga from SGU-Asia for adaptation and compression of central magic matrix developments based on W.Stanek's research and publication (refer list of publications).
Thanks of main author also to the SGU internship student Martin Eka Putra, Jemmi Dianto and assistant Florian Halfmann from University of Applied Sciences, Koblenz, testing the new 3D algorithm for arbitrary 64 cells magic cube with MATLAB®.

10. References

www.wolfram-stanek.de/stanek_sudoku_hoch_3_magisch_mindfestival2009.pdf
www.memomasters.de , "Mensch schlägt Computer"; MemoMasters 2009, 2008, Germany.
Spektrum der Wissenschaft, 2008-2,
Stanek Wolfram Prof, www.wolfram-stanek.de, 2010
http://en.wikipedia.org/wiki/Magic_square
http://www.multimagie.com/bi-magic squares

Magic Unit Checks for Physics and Extended Field Theory based on interdisciplinary Electrodynamics with Applications in Mechatronics and Automation

Prof. Dr.-Ing. Wolfram Stanek[1,2], Ir. Arko Djajadi, Ph.D[3]
and Edward Boris P Manurung, MEng[4]
[1]*University of Applied Sciences Koblenz*
[2]*Guest lecturer at Swiss German University, BSDCity-Jakarta*
[3]*Head of Research, in Mechatronics Department, Swiss German University, BSDCity-Jakarta*
[4]*Pro Rector of Swiss German University, BSDCity-Jakarta*
[1]*Germany*
[2,3,4]*Indonesia*

1. Introduction

What is the problem? The often recognised problem in mechatronics is a lack of experience in applying electrodynamic knowledge. Therefore a compact introduction in an extended Maxwell's field theory with interdisciplinary applications shall introduce a valuable key for all "Mechatronicists". All the described industrial developments were primarily based on electrodynamics, using innovative ideas, Maxwell's equations and both software and computer-aided simulation. However the focus of this publication is primarily on the advantage and necessity of electrodynamics inside mechatronics.
FIRST, the mighty capabilities of Unit Checks for deriving all central equations and formulas in physics are surprising and magic. These mostly unknown Unit Check methods will demonstrate the commonly unused fast derivation of famous and complex equations in physics from mechanics, electrodynamics up to quantum mechanics, Einstein's relativity formulas etc.
SECOND, the known Maxwell's equations in rest were extended and re-formulated for arbitrarily moving objects. Additionally, the sketched derivation of a unified equation for relativistic quantum electrodynamics based on Faraday and Einstein - including Maxwell's equations as a subset - will show further interdisciplinary applications in classic, quantum and relativistic physics.
THIRD, the structure identity of the complete eddy current equations in electrodynamics with respect to other disciplines in physics (i.e. hydrodynamics, thermodynamics, elastomechanics etc) opens a door for both quick analytical approximation and interdisciplinary development or optimisation of new mechatronic systems. Actual computer-based and analytical applications in the broad field of motor car production, robot gripper design, anti-vibration systems and complex hard disc drives will show the high efficiency and central position of extended Maxwell's equations in electrodynamics for automation and mechatronics.

This publication about interdisciplinary electrodynamics is based on research and development by Prof Stanek at University of Applied Sciences Koblenz (Germany), in addition to his guest lectures at Swiss German University SGU (BSDCity / Jakarta Indonesia) and Technical University Opole (Poland), his contributions at the REM conference Research and Education in Mechatronics (Stanek & Grueneberg, 2003), his own publications and his books about field theories and industrial mechatronics (Cassing & Stanek, 2002; Stanek et al., 2001), his results of an advised Master Thesis at SGU about robotics (Andries, 2003), his research and developments for motor car production (Stanek et al, 1984) and his own web sites about extended Electromagnetic Field Theory using Heaviside's streamlined re-design of Maxwell's equations and extensions (Stanek, 2010).

2. Electrodynamics as a central part in mechatronics

The fact that electrodynamics is a central part in mechatronics will be shown by different views of Maxwell's equations and interdisciplinary evaluations.

2.1 Electrodynamics based on Maxwell's equations

One of the most famous formulations in physics is the set of Maxwell's equations. Later, some basic equations will be shown or re-formulated and then extended.

2.1.1 Basic Maxwell's equations and constitutive relations

A compact overview of basic Maxwell's equations in differential and integral formulation with (nonlinear) constitutive relations is presented in this section.
Eq. (1) in Fig. 1 is Ampere-Maxwell's Law and eq. (2) Faraday-Lorentz' Law, both of which are called field equations. Eq. (3) is electric Gauss' Law and eq. (4) magnetic Gauss' Law, both are called source equations for Maxwell's field theory. **B** is magnetic flux density in **Vs/m²**, **H** is the magnetic field strength in **A/m**, **D** is displacement or electric flux density in **As/m²,** E is the electric field strength in **V/m, J** is the electric current density in **A/m², ρ** is the electric volume charge density in **As/m³,** Q is electric charge in **As**, and **∇** is the Nabla-Operator for vector analytical operations. For all bodies in rest, the dot (•) over **D** and **B** means partial derivatives of these characteristics with respect to time (here $d/dt=\partial/\partial t$). Simple mnemonics are shown in Fig.1.

Maxwell's equations in differential form

The basic set of Maxwell's equations (1) - (4) can be written in differential form:

$$\nabla \times \mathbf{H} = \mathbf{J} + \dot{\mathbf{D}} \quad (1) \qquad \nabla \times \mathbf{E} = -\dot{\mathbf{B}} \quad (2)$$

$$\nabla \cdot \mathbf{D} = \rho \quad (3) \qquad \nabla \cdot \mathbf{B} = 0 \quad (4)$$

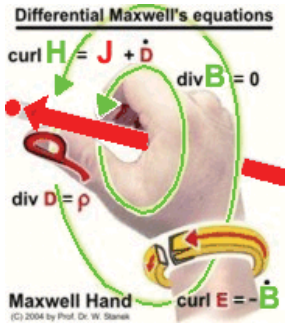


Fig. 1. Set of Maxwell's equations with equivalent mnemonics "Maxwell's Hand" (Stanek, 2002+2010)

Maxwell's equations in integral form

Using the known vector analysis laws by Stokes and Gauss we get from eq. (1) - (4):

$$\oint \mathbf{H}\,\mathrm{d}\mathbf{l} = \iint \mathbf{J}\,\mathrm{d}\mathbf{s} + \iint d\mathbf{D}/dt \cdot \mathrm{d}\mathbf{s} \quad (1a) \qquad \oint \mathbf{E}\,\mathbf{dl} = -\iint d\mathbf{B}/dt \cdot \mathrm{d}\mathbf{s} \quad (2a)$$

$$\oiint \mathbf{D}\,\mathrm{d}\mathbf{s} = \iiint \rho\, d\nu = Q \quad (3a) \qquad \oiint \mathbf{B}\,\mathrm{d}\mathbf{s} = 0 \quad (4a)$$

Because of primarily using the superior magnetic vector potential **A** shown in later equations, we introduce letter **s (=surface) for area**, **l is the length** and ν **(=nu) is the volume**. If we don't consider moving bodies, the terms d/dt are partial derivatives ∂/∂t.

Constitutive relations

The constitutive relations between the **classical field terms D, E, B, H and J**, also including both polarisations and external current sources, are defined by eq. (5) - (7):

$$\mathbf{D} = [\varepsilon]\,\mathbf{E} + \mathbf{P} \quad (5) \qquad \mathbf{B} = [\mu]\,\mathbf{H} + \mathbf{B}_p \quad (6) \qquad \mathbf{J} = [\gamma]\,\mathbf{E} + \mathbf{J}_e \quad (7)$$

Eq. (6) with details: $\mathbf{B} = \mu\,\mathbf{H} + \mathbf{B}_p = \mu_0\,\mathbf{H} + \mu_0\,\mathbf{M}_e + \mu_0\,\mathbf{M}_p = \mu_0\,(\mathbf{H} + \mathbf{M}_e) + \mu_0\,\mathbf{M}_p$ (6a)

In eq. (6a) $\mathbf{B}_p$ is the magnetic polarisation and $\mathbf{M}_p$ is the magnetisation in permanent magnets, $\mathbf{M}_e$ is the magnetisation in magnetic iron caused by an external field (index "e"), considering magnetic iron without permanent magnets $\mathbf{B}_p = 0$, without iron $\mathbf{M}_e = 0$, too (Oberretl, 2008). The material property $\mu = \mu_0 \cdot \mu_r$ is the permeability in ferromagnetic materials, $\varepsilon = \varepsilon_0 \cdot \varepsilon_r$ is the permittivity in dielectric materials and γ is the electrical conductivity. **P** is the electric polarisation, $\mathbf{J}_e$ are all possible external current sources. In most industrial applications magnetic material properties, primarily permeability, show non-linear characteristics, ref. Fig. 2. [$\mu_0 = 4\pi \cdot 10^{-7}$ Vs/Am = 1/ $(c^2 \cdot \varepsilon_0)$]

2.1.2 Extended Maxwell's equations considering moving bodies

The following four re-formulated Maxwell equations (1b) - (4b) can be used for all advanced calculations and computations in electrodynamics (with fields and waves), including constitutive relations [ref. to eq. (5)-(7)] and arbitrary movements of bodies (or particles) with **speed v**. The basis of these extensions is the relativity relation $(\mathbf{v}\cdot\nabla)\,\mathbf{A} = d\mathbf{A}/dt - \partial\mathbf{A}/\partial t$ (ref. to Einstein's Relativity Theory (Einstein, 1905), Helmholtz' theorems for moving objects (Cassing & Stanek, 2002; Stanek, 2010), and Sommerfeld's electrodynamics (Sommerfeld, 1988)), where **A** may be any vector, scalar or tensor. Furthermore these equations are the central basis for understanding interdisciplinary physics, especially structure identical formulations in i.e. hydrodynamics, diffusion, thermodynamics etc compared with directly derivable eddy current equations. Material properties of **[μ], [ε] and [γ]** in brackets shall be a reminder that they are often non-linear and additionally tensors.

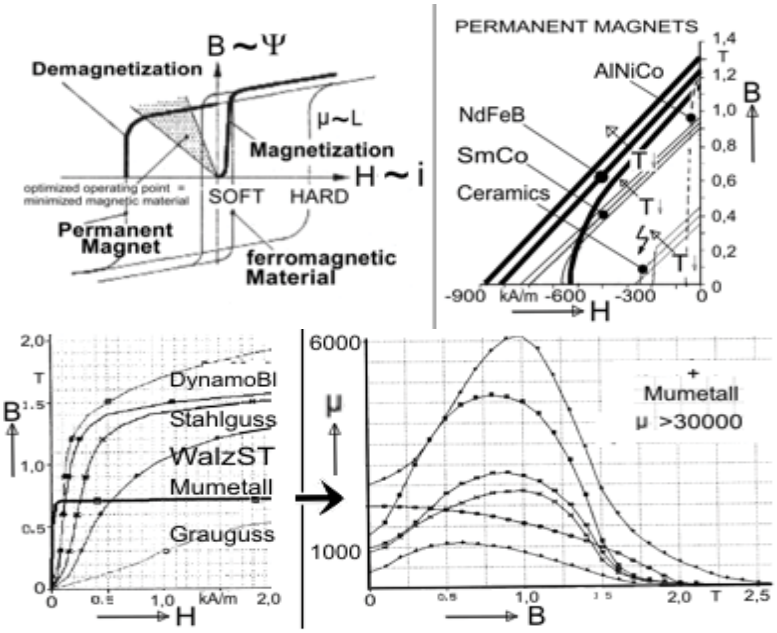


Fig. 2 Constitutive relations of permanent magnets + ferromagnetic materials (Cassing & Stanek, 2002; Stanek & Grueneberg, 2003)

1. extended Maxwell's equation **Ampere-Maxwell's Law**	$\nabla \times \mathbf{H}' - + \left(\frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla) \right) \cdot \mathbf{D} + \mathbf{J}$	(1b)
2. extended Maxwell's equation **Faraday-Lorentz' Law**	$\nabla \times \mathbf{E}' = - \left(\frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla) \right) \cdot \mathbf{B}$	(2b)
3. extended Maxwell's equation **electric Gauss' Law**	$\nabla \cdot \mathbf{D}' = \rho'$	(3b)
4. extended Maxwell's equation **magnetic Gauss' Law**	$\nabla \cdot \mathbf{B}' = 0'$	(4b)
→ **using B, H, D, E** etc *area* based vector analysis (8a)	$(\mathbf{v} \cdot \nabla) \cdot \mathbf{B} =$ $- curl\ (\mathbf{v} \times \mathbf{B}) + \mathbf{v} \cdot div\ \mathbf{B} - \mathbf{B} \cdot div\ \mathbf{v} + (\mathbf{B} \cdot grad) \cdot \mathbf{v}$	(8a)
→ using **A** (with **B** = curl **A**) *line* based vector analysis (8b)	$(\mathbf{v} \cdot \nabla) \cdot \mathbf{A} =$ $- \mathbf{v} \times curl\,\mathbf{A} + grad(\mathbf{v} \cdot \mathbf{A}) - (\mathbf{A} \cdot grad) \cdot \mathbf{v} - \mathbf{A} \times curl\,\mathbf{v}$	(8b)

Fig 3. Extended Maxwell's equations for moving bodies and basics in vector analysis (Cassing & Stanek, 2002; Stanek, 2010)

The transformation equations in *general* formulation are:

E' **=** E **+** v **x** B **+** ... ***further terms*** H' **=** H **-** v **x** D **+** ... ***further terms*** **-> refer to eq.** (8)

The additional field entities i.e. v x B and v x D - caused by moving bodies - are only 1 of 4 possible terms. The transformation equations in *simplified* formulation are therefore

$$\mathbf{E}' = \mathbf{E} + \mathbf{v} \times \mathbf{B} \qquad (2c) \qquad \text{and} \qquad \mathbf{H}' = \mathbf{H} - \mathbf{v} \times \mathbf{D} \qquad (1c)$$

and well known as basic Lorentz' Transformation. Using only this special transformation the transformed current caused by moved body (i.e. conductor) is

$$\mathbf{J}' = \mathbf{J} - \mathbf{v}\,\rho. \qquad (1e)$$

The following three examples shall deepen the background about influences of transformations:

1. EXAMPLE for evaluation with magnetic flux density terms. The derivation of Faraday-Lorentz' Law using equation (8a): Assuming special conditions/restrictions (in literature often not mentioned) i.e. incompressible materials div $\mathbf{v} = 0$, space independent constant movements ($\mathbf{B}$ grad) $\mathbf{v} = 0$ and in magnetic fields directly from magnetic Gauss' law always div $\mathbf{B} = 0$ the remaining term on the right side in eq. (8a) yields rot ($\mathbf{B} \times \mathbf{v}$) = - rot ($\mathbf{v} \times \mathbf{B}$) = - curl ($\mathbf{v} \times \mathbf{B}$). Inserting this result in Faraday's Law we can simply derive the extended 2. Maxwell's equation for moving bodies:
differential Faraday - Lorentz' – Law

$$\text{curl}\,\mathbf{E}' = -\,d\,\mathbf{B} / dt = -\,\partial\,\mathbf{B} / \partial\,t + \text{curl}\,(\mathbf{v} \times \mathbf{B}) \qquad (2d)$$

Using equation (8b) with the same conditions mentioned above, we get eq. (2d) with $\nabla \times \mathbf{A}$ = Nabla x $\mathbf{A}$ = curl $\mathbf{A}$ = $\mathbf{B}$. The first term on the right side of this equation (2d) was proved by Faraday, the second one by Lorentz. NOTE: using this vector analytical formulation we get the Lorentz-Term $\mathbf{E} = \mathbf{v} \times \mathbf{B}$ automatically! The famous Lorentz law is therefore a (very important) vector identity, but not really a separate physical law.

2. EXAMPLE for evaluation with electric flux density terms. The derivation of Ampere-Maxwell's Law using equation (8a): Assuming special conditions/restrictions as in the Example 1 (i.e. incompressible materials div $\mathbf{v} = 0$, space independent constant movements ($\mathbf{D}$ grad) $\mathbf{v} = 0$ and in electric fields directly from electric Gauss' law div $\mathbf{D} = \nabla \bullet \mathbf{D}$ the remaining term on the right side in eq. (8a) yields for the cross product rot ($\mathbf{D} \times \mathbf{v}$) = - rot ($\mathbf{v} \times \mathbf{D}$) = - curl ($\mathbf{v} \times \mathbf{D}$) and in opposite to the Faraday-Lorentz' Law in eq. (2d) for the Ampere-Maxwell's Law in equation (1d) an additional term. Considering simplified conditions like non-relativistic, linear and constant Movements yields eq. (3). But generally $\mathbf{v} \bullet \text{div}\,\mathbf{D}' = \mathbf{v} \bullet \boldsymbol{\rho}'$ is valid, ref. to eq. (3b) (Sommerfeld, 1988; Stanek, 2010).
Inserting these results in Ampere-Maxwell's Law eq. (1a) we can derive the extended eq. (1b). Maxwell's equation for moved bodies or particles with the following expressions:

$$\text{curl}\,\mathbf{H}' = \mathbf{J} + d\,\mathbf{D} / dt = \mathbf{J} + \partial\,\mathbf{D} / \partial\,t + \mathbf{v} \bullet \rho - \text{curl}\,(\mathbf{v} \times \mathbf{D}) \qquad (1d)$$

The first term on the right side of this equation (1d) was proved by Ampere, the third term by Rowland, the second term by Hertz (suggested and introduced by Maxwell), and the fourth term by Roentgen. NOTE: using this vector analytical formulation we get the "dualism" of the Lorentz-Term $\mathbf{H} = -\,\mathbf{v} \times \mathbf{D}$ automatically! The Rowland and Roentgen terms are therefore (important) vector identities, but not really separate physical laws.

3. EXAMPLE for proof of extended eq. (1b) and eq. (2b) Maxwell's equations using the famous HELMHOLTZ' formula. Helmholtz derived for any arbitrary vector flux $\mathbf{X}$ in physics (i.e. hydrodynamics) through a moved ($\mathbf{v}$) and simultaneously deformable area element in his curl laws - as a subset of (8a) –which yields the following formula:

$$d\,\mathbf{X} / dt = \partial\,\mathbf{X} / \partial t + \text{curl}\,(\mathbf{X} \times \mathbf{v}) + \mathbf{v}\,\text{div}\,\mathbf{X} \quad (*)$$

Inserting this Helmholtz' formula (*) in the Maxwell equations (1a) and (2a) – with the prerequisite of the same above mentioned conditions and **X** = **B** alternatively **X** = **D** - we immediately get the extended Maxwell's equations (1b) and (2b) in the 1. and 2. example! NOTE: using (*) the extended Maxwell's equations are derivable without any knowledge in vector analysis. The Helmholtz' formula is ingenious and the basis for Lorentz, Minkowski and Einstein, too. Helmholtz derived his formula visualising - like a "mnemonics artist" - moved and deformable geometric elements. Nevertheless Helmholtz' formula d**X** /dt neglects the LAST term, here (**X** ∇) **v** inside (**v** ∇) **X** (i.e. additional rotations), refer to (8a) and (8b) !

2.1.3 Extended Maxwell's equations in 4-dimensional formulation

Another compact expression of Maxwell's equations (i.e. in vacuum without materials and no movable bodies) can be derived, using 4-dimensional expressions (Sommerfeld, 1988; Cassing & Stanek, 2002)**:**

1. space-time operator □ (x, y, z, i · c · t) with d'Alembert $\Box \equiv \Delta - 1/c^2 \cdot \partial^2/\partial t^2 = \sum_{i=1}^{4} \partial^2/\partial x_i^2$
2. **A**-φ-Potential **Ω** (A_x, A_y, A_z, i · φ / c) with c = $1/\sqrt{(\varepsilon_0 \cdot \mu_0)}$, i = $\sqrt{(-1)}$ and
3. current densities **Γ** (J_x, J_y, J_z, i · ρ · c) respectively **J′** = **v** · ρ with condensed results:

$$\text{a) } \Box\,\mathbf{\Omega} = -\mu_0 \cdot \mathbf{\Gamma}\,, \quad \text{b) } \nabla \cdot \mathbf{\Omega} = 0\,, \quad \text{c) } \nabla \cdot \mathbf{\Gamma} = 0\,, \quad \text{d) } \mathbf{F} = \mu_0 \cdot \mathbf{G} = \nabla \times \mathbf{\Omega} \qquad (9)$$

where **F** (**B**, -i·**E**/c) and **G** (**H**, -i·c**D**) define the electromagnetic Maxwell field tensors.

2.1.4 Extended Maxwell's equations in quantum electrodynamics

Quantum electrodynamics is a complex interdisciplinary field, but is not normally used daily by practical mechatronics engineers. On the other side many phenomena (duality of wave and particle, tunnel diode, special superconductivity up to quantum computers etc) are important and must be handled with a background of this superior theory based on the integration of electrodynamics, quantum mechanics and (for relativistic processes) relativity theory (Cassing & Stanek, 2002). As a compromise only the resulting extended Maxwell equations in quantum electrodynamics will be shown in (1f) - (4f).

$$\nabla \times \mathbf{H} = \mathbf{J} + \dot{\mathbf{D}} - \kappa^2 \cdot \mathbf{A} / \mu_0 \qquad (1f)$$

$$\nabla \times \mathbf{E} = -\dot{\mathbf{B}} \qquad (2f)$$

$$\nabla \cdot \mathbf{D} = \rho - \kappa^2 \cdot \varphi \cdot \varepsilon_0 \qquad (3f)$$

$$\nabla \cdot \mathbf{B} = 0 \qquad (4f)$$

These extended Maxwell's equations, which are called Proca's equations, additionally describe special phenomena in quantum electrodynamics (Lehner, 1994; Cassing&Stanek, 2002). These further quantum terms consist of classical magnetic vector potential **A**, electrical scalar potential φ, the special term $\kappa^2 = (m_0 \cdot c / \hbar)^2$ and material properties in vacuum (namely permeability μ_0 and permittivity ε_0).

The term κ^2 is famous in quantum mechanics, because κ is Compton's frequency divided by the speed of light c or Einstein's energy in view of quantum mechanics. The mass in rest is 0, the universal Planck's constant in quantum mechanics is $\hbar$ (= h / $2\pi \approx 1 \cdot 10^{-34}$ J s).

2.2. Interdisciplinary evaluation of Maxwell's equations

From Maxwell's equations we can directly derive all central relations for electromagnetic waves and fields, eddy current equations, structure identities inside electrodynamics and with other physical disciplines as well. Ref. to all possible derivations in chap. 2.2.4.

2.2.1 Electromagnetic field and wave equations

Electrodynamics as one compact equation including polarisations and movable bodies is given by:

$$curl\frac{1}{\mu}curl\mathbf{A} = \mathbf{J}_e + curl\frac{1}{\mu}\mathbf{M}_p + \frac{\partial \mathbf{P}}{\partial t} + \mathbf{v}\cdot\rho + \left(\gamma + \varepsilon\frac{\partial}{\partial t}\right)\cdot\left[-grad\varphi - \frac{\partial \mathbf{A}}{\partial t} + \mathbf{v}\times curl\mathbf{A}\right] \tag{10}$$

Choice of gauges in electrodynamics is important for evaluation of fields and waves, because potentials **A** and φ are not unique (Ψ scalar magnetic potential), eq.(10).

$$\mathbf{A} = \mathbf{A}^* - \nabla\psi\,, \quad \varphi = \varphi^* + \partial\psi / \partial t \tag{10a+b}$$

$$\Delta\mathbf{A} - \mu\varepsilon\frac{\partial^2\mathbf{A}}{\partial t^2} - \mu\gamma\frac{\partial\mathbf{A}}{\partial t} = \nabla\cdot\left[\nabla\mathbf{A} + \mu\varepsilon\frac{\partial\varphi}{\partial t} + \mu\gamma\varphi\right] \tag{11}$$

The most used gauges are the complete Lorentz gauge […]=0, eq. (11) and reduced Lorentz gauge $\nabla\mathbf{A} = -\,\mu\varepsilon\cdot\partial\varphi/\partial t$ for waves, and Coulomb gauge $\nabla\mathbf{A}=0$ for eddy current and static applications. Wave equations from eq. (11) using eq. (3) and polarisations:

$$\Delta\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu_0\left(\mathbf{J} + \nabla\times\mathbf{M}_p + \frac{\partial\,\mathbf{P}}{\partial t}\right) \tag{12a}$$

$$\Delta\phi_S - \frac{1}{c^2}\frac{\partial^2\varphi_S}{\partial t^2} = -\frac{1}{\varepsilon_0}\left(\rho - \nabla\cdot\mathbf{P}\right) \tag{12b}$$

Wave equations derived from concentrated field elements in electric circuits with resistance R, conductance G, capacitance C, inductance L (mutual inductance M) yield the same result for voltage V and current I, instead of **A** or φ, as shown in eq.(11) respectively eq.(12a,b).

2.2.2 Eddy current equation in electrodynamics

With $(\varepsilon\,\partial/\partial t) = 0$, eq.(10) leads to interdisciplinary usage of eddy current equation(13).

$$curl\ \frac{1}{\mu}curl\ \mathbf{A} = \mathbf{J} - \gamma\cdot grad\ \varphi + curl\ \frac{1}{\mu}\mathbf{M}_p - \gamma\cdot\frac{\partial\mathbf{A}}{\partial t} + \gamma\cdot\mathbf{v}\times curl\ \mathbf{A} \tag{13}$$

The current density J includes all further electrical excitations shown in eq. (10).

2.2.3 Static equations inside electrodynamics with identical structure

From Maxwell's equations we get formulations with identical structure for magnetic fields in magnetostatics, electric fields in electrostatics and electric current flow. In Fig. 4 six identical fields are sketched for different areas inside electrodynamics. The field map for only electrostatics automatically yields the results for the other shown disciplines, refer to eq.(19a, 20a). The field maps were evaluated for the centre of the applications shown, while neglecting the leakage fluxes i.e. of capacitor and current sheets.

"Trial and Error" field mapping proved by field numerical computations with FEM program MagnetoCAD is shown in Fig.4. Field mapping rules in Fig. 4a considered field lines and equipotential lines as perpendicular, equidistantly arranged and sketched by means of curvilinear squares.

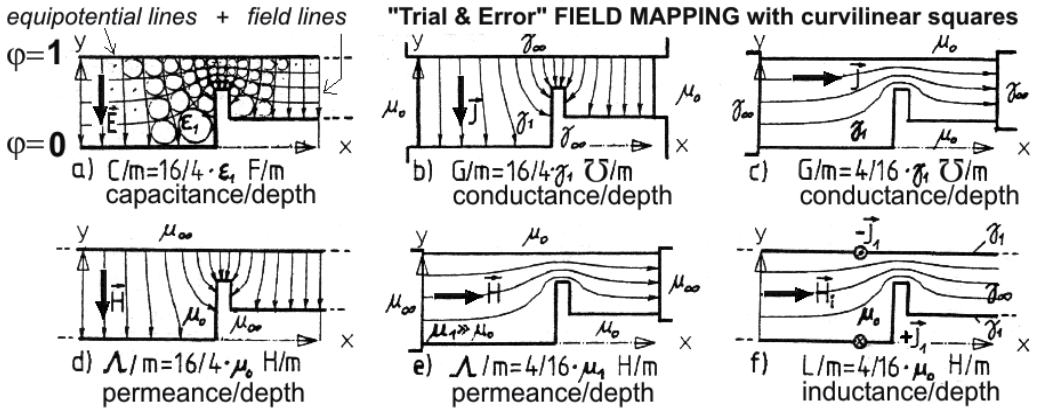


Fig. 4. Application of one field map to six interdisciplinary cases inside electrodynamics (Stanek, 2002)

2.2.4 All possible Mind Map derivations from variations of "Maxwell's Hand"

All central derivations from Maxwell's equations with respect to all important phenomena inside electrodynamics are developed by the author and visualised as a new Mind Map with 10 memorable Memo Maps. These maps are based on variations of Maxwell's "Right Hand Rule" and Brain Power Rules (Stanek et al, 2006). Starting from differential equations we can formulate all central equations governing electrodynamics and interdisciplinary physics. These Memo Maps are valuable mnemonics for necessary derivations, useful backgrounds and compact results. Memorising these pictures is easy for us to bear all derivations in mind concerning the variety of extended Maxwell's equations.

The Mind Map can be found on a special web site prepared by the author (Stanek, 2010) as given in Fig. 5. Most of all these formulas and equations can be derived using the powerful unit check method shown in the next chapter 2.2.5.

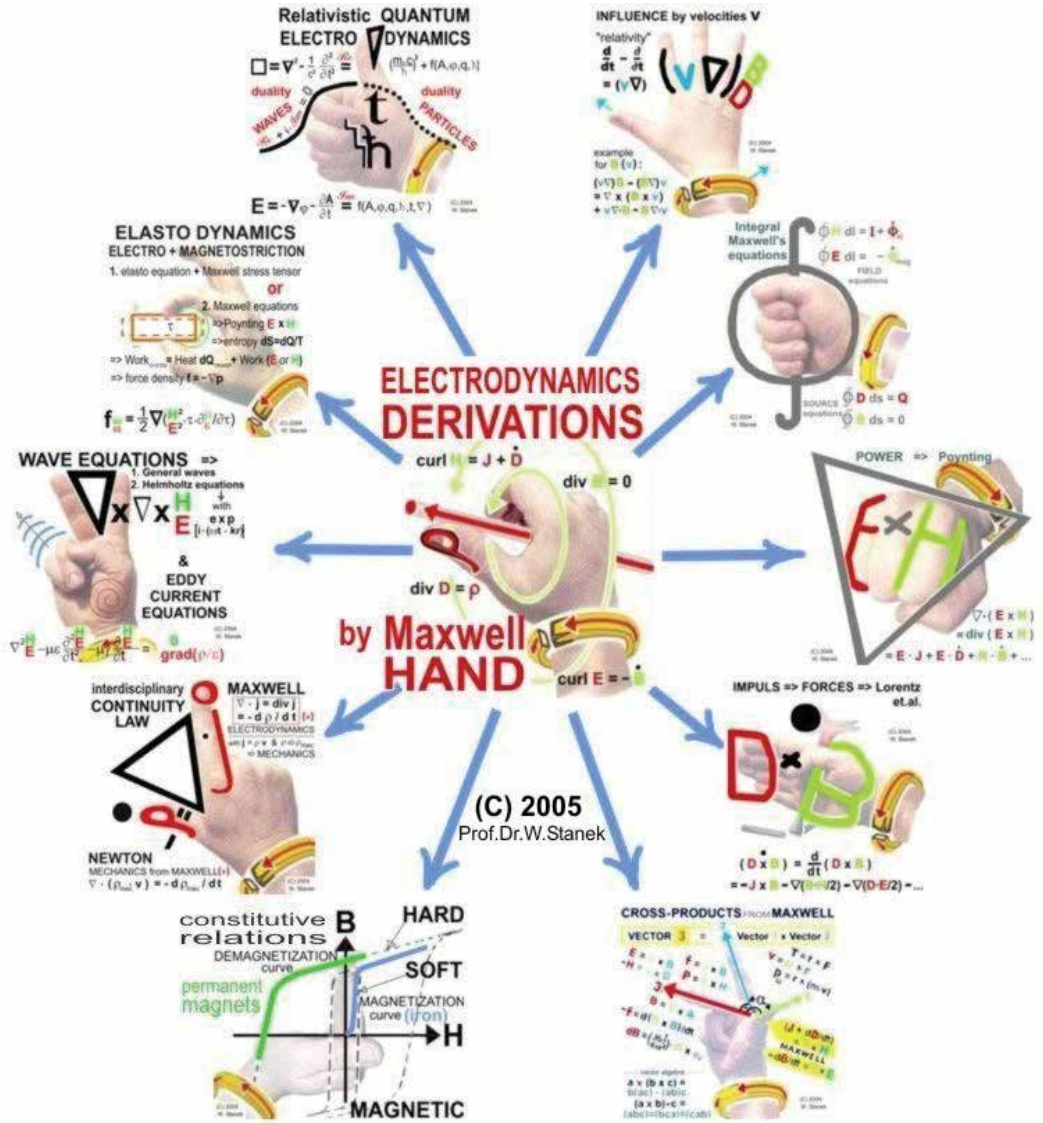


Fig. 5. Mind Map for central derivations from Maxwell's equations (Stanek, 2010)

2.2.5 The mighty method in physics: Deriving equations by unit checks

Following questions (Q1 … Q20) and answers (A x.y) will demonstrate and train this useful method using both sides of our brain to understand, to derive, to learn and to recall most of the important formulas in physics without any effort.

Q1) *What will happen with an electric Charge Q placed in an electric field* ***E****?*	
The scalar Q in As and vector **E** in V/m build the product $Q \cdot \mathbf{E}$.	(A1.1)
Equivalent unit equation: $As \cdot V/m = Ws/m = Nm/m = N$ = Newton → Force $\mathbf{F_{el}}$	(A1.2)
The result is Coulomb's law, an electric force $\mathbf{F_{el}} = Q \cdot \mathbf{E}$	(A1.3)
Q2) *What is equivalent to a space-depending electric potential* φ *defined by gradient "grad"?*	
This term can be written as a product of Nabla operator and electric potential $\nabla\varphi$	(A2.1)
Equivalent unit equation: $1/m \cdot V = V/m$ electric field strength **E**	(A2.2)
Regarding signs for "grad" in mathematics and **E** in physics the result is: $\mathbf{E} = -$ grad φ	(A2.3)
Q3) *What will happen in a magnetic field* ***B*** *moving a particle / body with a uniform speed* ***v****?*	
Both vectors **v** in m/s and **B** in Vs/m^2 build a cross product $\mathbf{v} \times \mathbf{B}$.	(A3.1)
Equivalent unit equation: $m/s \times Vs/m^2 = V/m$ → electric field strength **E**	(A3.2)
The result is the additionally induced *electric* Lorentz' field strength $\mathbf{E_L} = \mathbf{v} \times \mathbf{B}$	(A3.3)
Q4) *What will happen in an electric field* ***D*** *moving a particle / body with a uniform speed* ***v****?*	
Both vectors **v** in m/s and **D** in As/m^2 build a cross product $\mathbf{v} \times \mathbf{D}$ → (ref. to Q18 !)	(A4.1)
Equivalent unit equation: $m/s \times As/m^2 = A/m$ → magnetic field strength **H**	(A4.2)
The result is the additional *magnetic* Lorentz' field strength $\mathbf{H} = \mathbf{v} \times \mathbf{D} = -\mathbf{D} \times \mathbf{v}$	(A4.3)
Applying "∇" operator on **H** the result is Roentgen's current $\mathbf{J_{Roe}} = \nabla \times (\mathbf{D} \times \mathbf{v}) =$ curl $(\mathbf{D} \times \mathbf{v})$	(A4.4)
Q5) *What is equivalent to an electric charge density* ρ *moved with the speed* **v**?	
The scalar ρ in As/m^3 and vector **v** in m/s build the product $\rho \cdot \mathbf{v}$.	(A5.1)
Equivalent unit equation: $As/m^3 \cdot m/s = A/m^2$ additional electric current density $\mathbf{J_{Row}}$	(A5.2)
The result is Rowland's current density $\mathbf{J_{Row}} = \rho \cdot \mathbf{v}$.	(A5.3)
Q6) *How much is the force on a current carrying conductor or moved* ρ *in a magnetic field* **B**?	
The physical entities ρ, **v** and **B** build the cross product $\rho \cdot \mathbf{v} \times \mathbf{B}$.	(A6.1)
Equivalent unit equation: $As/m^3 \cdot m/s \times Vs/m^2 = Ws/m^4 = Nm/m^4 = N/m^3$ → force density **f**	(A6.2)
The result is Lorentz' force density caused by electric currents $\mathbf{f_L} = \mathbf{J} \times \mathbf{B}$	(A6.3)
Q7) *What will happen when a magnetic flux density* ***B*** *is time-changing through a loop?*	
The action $\partial\mathbf{B} / \partial t$ causes a reaction in a loop which must be a negatively signed vector, too.	(A7.1)
Equivalent unit equation: $1/s \cdot Vs/m^2 = 1/m \cdot V/m$ → ∇ applied on electric field strength **E**	(A7.2)
The result is Faraday's law or Maxwell's second (field) equation $-\partial\mathbf{B} / \partial t = \nabla \times \mathbf{E}$ = curl **E**	(A7.3)
Q8) *Which physical entity will be produced by an electric current density* ***J****?*	
All currents will produce a magnetic field strength **H** easily derived by following unit check:	
Equivalent unit equation: $A/m^2 = 1/m \cdot A/m$ → ∇ applied on magnetic field strength **H**	(A8.1)
The result is basic Ampère's law or Maxwell's first (field) equation $\mathbf{J} = \nabla \times \mathbf{H}$ = curl **H**	(A8.2)
Q9) *Which source divergence "div" of a physical entity produces electric charge density* ρ?	
This relation can be written as a product of Nabla operator and electric potential $\nabla \cdot$ "?" $= \rho$	(A9.1)
Equivalent unit equation: $1/m \cdot$ '?' $= As/m^3$ or '?' $= As/m^2$ electric flux density **D**	(A9.2)
The result is electric Gauss' law or Maxwell's third (source) equation $\nabla \cdot \mathbf{D}$ = div $\mathbf{D} = \rho$	(A9.3)
Q10) *Which magnetic source divergence "div" of a physical entity is always zero?*	
This relation can be written as a product of Nabla operator and electric potential $\nabla \cdot$ "?" $= 0$	(A10.1)
Equivalent unit equation: $1/m \cdot$ '?' $= Vs/m^3$ (fictive monopole) → magnetic flux density **B**	(A10.2)
The result is magnetic Gauss' law or Maxwell's fourth (source) equation $\nabla \cdot \mathbf{B}$ = div $\mathbf{B} = 0$	(A10.3)

Q11) *Which time-changing physical entity* ***X*** *will produce a current density* ***J*** *in air (vacuum)?*	
This relation can be written as $\partial \mathbf{X} / \partial t = \mathbf{J}$, where **X** is the searched unknown.	(A11.1)
Equivalent unit equation: 1/s · '?' = A/m² or '?' = As/m² → electric flux density **D**	(A11.2)
The result is Maxwell's displacement current $\mathbf{J_D} = \partial \mathbf{X} / \partial t$ (= 2nd part of 1. Maxwell's equation)	(A11.3)
Q12) *What is equivalent to the source "div" of a moved charge density* ρ with the speed **v**?	
Applying Helmholtz' law div (curl **H**) = 0 on eq. (A8.2) with (A11.2) or from $\nabla \cdot (\rho \cdot \mathbf{v}) =$ "?"	(A12.1)
Equivalent unit equation: 1/m · (As/m³ · m/s) = A/m² = 1/s · As/m³ → $\partial / \partial t \cdot$ charge density ρ	(A12.2)
The result is Maxwell's continuity law in electrodynamics $\nabla \cdot (\rho \cdot \mathbf{v}) = \nabla \cdot \mathbf{J} = - \partial \rho / \partial t$	(A12.3)
Q13) *What is equivalent to the source "div" of a moved mass density* ρ_M with the speed **v**?	
From $\nabla \cdot (\rho_M \cdot \mathbf{v}) =$ "?" unit equation: 1/m · (kg/m³ · m/s) = 1/s·kg/m³ → $\partial / \partial t \cdot$ mass density ρ_M	(A13.1)
The result is Newton's continuity law in mechanics $\nabla \cdot (\rho_M \cdot \mathbf{v}) = \nabla \cdot \mathbf{J} = - \partial \rho_M / \partial t$	(A13.2)
Q14) *What is equivalent to the mechanical (im)pulse density* $\mathbf{p_M} = \rho_M \cdot \mathbf{v}$ *in electrodynamics?*	
Equivalent unit equation: kg/m³ · m/s = kg · m/s² · s/m³ = Ns/m³ = Ws/m³ · s/m = As/m3 ·Vs/m	(A14.1)
The unit Vs/m defines the magnetic vector potential **A** with its subset **B** = curl **A** = $\nabla \times \mathbf{A}$.	(A14.2)
Moving like Maxwell's hand for **J** = curl **H** we see that 2-D fields can be calculated by 1-D !	(A14.3)
The result is the equivalence of $\mathbf{p_M} = \rho_M \cdot \mathbf{v}$ in mechanics and $\mathbf{p_{EM}} = \rho \cdot \mathbf{A}$ in electrodynamics	(A14.4)
Q15) *What is the relation between (im)pulse density* $\mathbf{p_M}$ *or* $\mathbf{p_{EM}}$ *and force density* **f**?	
Equivalent unit equation from (A14.1): Ns/m³ · '?' = N/m³ or '?' = 1 / s → time operator d/dt	(A15.1)
a) The result in mechanics is Newton's force law $\mathbf{f_M} = d(\rho_M \cdot \mathbf{v}) / dt = \mathbf{v} \cdot \partial \rho_M / \partial t + \rho_M \cdot \partial \mathbf{v} / \partial t$	(A15.2)
b1) Result in electrodynamics is Lorentz' force law $- \mathbf{f_{EM}} = d(\rho \cdot \mathbf{A}) / dt = \mathbf{A} \cdot \partial \rho / \partial t + \rho \cdot \partial \mathbf{A} / \partial t$	(A15.3)
With re-formulated unit equation for last term: Vs/m · As/m³ / s = As/m³ · 1/m · V → $\rho \cdot \nabla \varphi$	(A15.4)
b2) Result in electrodynamics is Lorentz' force law $\mathbf{f_{EM}} = d(\rho \cdot \mathbf{A}) / dt = - \rho \cdot \nabla \varphi - \rho \cdot \partial \mathbf{A} / \partial t$	(A15.5)
Q16) *What is the relation between interdisciplinary force* **F** *and (potential) energy* W?	
From N = '?' · Nm or '?' = 1 / m → space operator ∇, with **P** = pulse, W = energy, follows:	(A16.1)
d'Alembert's force $\mathbf{F} = - \nabla W = - d W / d\mathbf{r} = -d W / dt \cdot dt / d\mathbf{r} = d\mathbf{P} / dt \rightarrow d(\mathbf{P} \cdot \mathbf{v} + W_{pot}) / dt = 0$	(A16.2)
Integration yields the non-relativistic energy law in classic physics: $W_{total} = W_{kin} + W_{pot} =$ const	(A16.3)
Q17) *a) What is the relation between energy W of an electromagnetic wave and frequency* ν	
Equivalent unit equation: Ws = '?' · 1/s or '?' = Ws · s → Planck's constant h (or $\hbar$)	(A17.1)
Result is Planck's energy relation: $W = h \cdot \nu$ or $W = \hbar \cdot \omega$ or complex → $\underline{W} = \mathbf{i} \cdot \hbar \cdot \partial / \partial t$	(A17.2)
b) What is the relation between pulse **P** of an electromagnetic wave and its wave length $\boldsymbol{\lambda}$?	
Equivalent unit equation: Ns = Nm s · 1/m = Ws · s · 1/m with wave vector **k** yields both the	(A17.3)
→ de Broglie's pulse: $\mathbf{P} = h / \boldsymbol{\lambda}$ or $\mathbf{P} = \hbar \cdot \mathbf{k}$ or complex with $\mathbf{i} = \sqrt{(-1)}$ → $\underline{\mathbf{P}} = - \mathbf{i} \cdot \hbar \cdot \nabla$	(A17.4)
→ de Broglie's wave equation: $\Psi(\mathbf{r}, t) = \Psi_0 \cdot \exp[\, \mathbf{i} \cdot (\omega \cdot t - \mathbf{k} \cdot \mathbf{r})\,] = \Psi_0 \cdot \exp[\, \mathbf{i} / \hbar \cdot (W \cdot t - \mathbf{P} \cdot \mathbf{k})\,]$	(A17.5)

c) Can we derive the relation for the uncertainty principle in quantum mechanics where neither energy W & time t nor pulse **P** & place **r** can be exactly determined at the same time?

From (A17.2) & (A17.4) we get Planck's constant $h = W / \nu = \mathbf{P} \cdot \boldsymbol{\lambda}$, where $\nu \sim 1/t$ & $\boldsymbol{\lambda} \sim \mathbf{r}$. (A17.6)

and directly in "small Δ - view" → Heisenberg's uncertainty relation $h \approx \Delta\mathbf{P} \cdot \Delta\mathbf{r} \approx \Delta W \cdot \Delta t$ (A17.7)

d) Inserting (A17.2) & (A17.4) in (A16.3) with mass M and kinetic energy $W_{kin} = M \cdot \mathbf{v}^2 / 2$ (A17.8)

→ non-relativistic Schrödinger's equation: $\hbar^2 / (2 \cdot M) + W_{pot} = \mathbf{i} \cdot \hbar \cdot \partial / \partial t$ [applied on $\Psi(r, t)$] (A17.9)

Q18) *How can we easily derive a formula for arbitrarily formed current loops?*

Considering the electric Gauss' law we quickly derive $d\mathbf{D} = dQ / (4\pi r^2) \cdot \mathbf{e}_r$, producing a (A18.1)

magnetic field $d\mathbf{H}$ by moved charges $\mathbf{v} \cdot dQ$ in units: $m/s \cdot As = A \cdot m \rightarrow \mathbf{v}_{dl} \cdot dQ = I \cdot \mathbf{dl}$ (A18.2)

With unit equation: $m/s \cdot As/m^2 = A/m$ the transformation from $d\mathbf{D}$ to $d\mathbf{H}$ is $\mathbf{v}_{dl} \times d\mathbf{D}$ (= A4.3) (A18.3)

The result is Biot-Savart's law: $d\mathbf{H} = \mathbf{v}_{dl} \times d\mathbf{D} = \mathbf{v}_{dl} \cdot dQ / (4\pi\mathbf{r}^2) \times \mathbf{e}_r = I / 4\pi \cdot (\mathbf{dl} \times \mathbf{e}_r) / \mathbf{r}^2$ (A18.4)

Q19) *What is the difference between total derivative "d / dt" and partial derivative "∂ / ∂t" ?*

i.e. $d\mathbf{B}(t, r) / dt = \partial\mathbf{B} / \partial t + \partial\mathbf{B} / \partial\mathbf{r} \cdot \partial\mathbf{r} / \partial t = \partial\mathbf{B} / \partial t + \partial\mathbf{B} / \partial\mathbf{r} \cdot \mathbf{v} = \partial\mathbf{B} / \partial t + (\mathbf{v}\,\nabla)\,\mathbf{B}$ (A19.1)

Basis for viewing point of moved systems → Einstein's special relativity theory (ref. Q20) (A19.2)

Vector gradient $(\mathbf{v}\,\nabla)\,\mathbf{B} = d\mathbf{B} / dt - \partial\mathbf{B} / \partial t$ implies additional fields in Lorentz' transformation (A19.3)

Applying unit checks i.e. $\mathbf{B} = f$ (temperature T) yields → additional fields $\partial\mathbf{B} / \partial T \cdot \partial T / \partial t$ etc (A19.4)

Most simple transformation from i.e. Faraday's law: $-\,d\mathbf{B}\,(\mathbf{v}) / dt = \text{curl}\,\mathbf{E}' = \text{curl}\,(\mathbf{E} + \mathbf{v} \times \mathbf{B})$ (A19.5)

Q20) *Which unit re-design leads from classic physics to Einstein's etc most famous formulas?*

Newton's force law in mechanics $\mathbf{f}_M = d(\rho_M \cdot \mathbf{v}) / dt$ (ref. Q15a) → $\mathbf{F}_M = d(M \cdot \mathbf{v}) / dt$ (M is mass) (A20.1)

Equivalent unit equation: $1/s \cdot kg \cdot m/s = kg \cdot m/s^2 = N = Nm / m = Ws / m \rightarrow Ws = kg \cdot m^2/s^2$ (A20.2)

The result is Einstein's equation: $\underline{E = M \cdot c^2}$ equivalence of energy E and mass M, where (A20.3)

the velocity unit m^2/s^2 means the square of the speed of light $c = 1 / \sqrt{(\mu_0 \cdot \varepsilon_0)} = \text{const.}$ (A20.4)

Viewing (ζ) moved body (Δt', v), time in rest Δt → Pythagoras yields: $\underline{(c \cdot \Delta t')^2 = (v \cdot \Delta t')^2 + (c \cdot \Delta t)^2}$ (A20.5)

or Einstein's time dilatation: $\Delta t' = \Delta t / \sqrt{[\,1 - (v/c)^2\,]}$. Extending (A20.5) by unit checks to the (A20.6)

relativistic energy equation: $E^2 = W_{tot}^2 = W_{kin}^2 + W_{pot}^2$, with pulse $\mathbf{P} = M \cdot \mathbf{v}$ & mass in rest M_0 (A20.7)

$(M \cdot c^2)^2 = (P \cdot c)^2 + (M_0 \cdot c^2)^2$ or Lorentz-Einstein's mass equation: $M = M_0 / \sqrt{[\,1 - (v/c)^2\,]}$ (A20.8)

Complex notation of (A20.8) using (A17.2) & (A17.4) leads to Klein-Gordon's equation or

→ relativistic Schrödinger's equation: $\Box = \nabla^2 - 1/c^2 \cdot \partial^2/\partial t^2 = (M_0 \cdot c / \hbar)^2$ ("space-time operator") (A20.9)

2.2.6 Electrodynamics compared with other physical disciplines

Central physical disciplines are compared with electrodynamics, neglecting Maxwell's $d\mathbf{D}/dt$ – term in eq. (14) - (16), (18a), (19a), (20a) and considering it in eq. (17) - (20).

Analogous expressions can be derived for diffusion equation in chemistry, Newton's mechanics, optics and acoustics etc. In eq. (14) the curl M_p-term is re-formulated as grad M_i - Σ term of M_p-components in i = x, y. Eq. (15) is also central for applications in aerodynamics. All these field equations (14) - (16) in Cartesian 2-dimensional coordinates will show identical structure, refer to Fig. 7. The central equation in non-linear elastodynamics is given by eq. (17), where μ_m and λ_m are Lamé characteristics for material elasticity, σ for non-linear tensions, u for mechanical displacement and f for external forces. Assuming linearity ($\text{div}\sigma = 0$) and incompressible media ($\text{div}\,u = 0$), elastodynamics is based on *wave equations* (18) - (20) with identical structure.

Electrodynamics Maxwell etc	$curl\,(1/\mu)\,curl\,\mathbf{A}$	$=\mathbf{J}-grad\left[\gamma\cdot\varphi\pm(1/\mu_i)\cdot M_i\right]$	$-\gamma\cdot\frac{\partial\mathbf{A}}{\partial t}+$	$\gamma\cdot\mathbf{v}\times curl\,\mathbf{A}$	(14)
Hydrodynamics Navier-Stokes etc	$curl\eta\,curl\,\mathbf{v}$	$=\mathbf{f}-grad\left[p+\rho_m\cdot\mathbf{v}^2/2\right]$	$-\rho_m\cdot\frac{\partial\mathbf{v}}{\partial t}+$	$\rho_m\cdot\mathbf{v}\times curl\,\mathbf{v}$	(15)
Thermodynamics Fourier-Helmholtz	$div\,\lambda\,grad\,T$	$=Q-grad\,q_s$	$-c_p\rho_m\cdot\frac{\partial T}{\partial t}+$	$c_p\rho_m\cdot\mathbf{v}\cdot grad\,T$	(16)
Elastodynamics Newton-Euler,	$\mu_m\cdot\Delta\mathbf{u}-\rho_m\cdot\partial^2\mathbf{u}/\partial t^2=-\mathbf{f}-(\mu_m+\lambda_m)\cdot grad\,div\,\mathbf{u}-div\,\boldsymbol{\sigma}$				(17)

Lagrange etc	$\Delta\mathbf{u}-(\rho_m/\mu_m)\cdot\partial^2\mathbf{u}/\partial t^2=-1/\mu_m\cdot\mathbf{f}$	(18)	Elastostatics	$\Delta\mathbf{u}=-1/\mu_m\cdot\mathbf{f}$	(18a)
Electrodynamics Maxwell, Ampere, Faraday,Gauss etc	$\Delta\mathbf{A}-(1/c^2)\cdot\partial^2\mathbf{A}/\partial t^2=-\mu_0\cdot\mathbf{J}$	(19)	Magnetostatics	$\Delta\mathbf{A}=-\mu_0\cdot\mathbf{J}$	(19a)
	$\Delta\varphi-(1/c^2)\cdot\partial^2\varphi/\partial t^2=-1/\varepsilon_0\cdot\rho$	(20)	Electrostatics	$\Delta\varphi=-1/\varepsilon_0\cdot\rho$	(20a)

Fig. 6. Interdisciplinary vector analytical structure identities in physics (Cassing & Stanek, 2002; Bronstein, 1995)

2.2.7 Electrodynamics directly integrated with other physical disciplines

Magnetohydrodynamics: Hydrodynamics + Electrodynamics + Thermodynamics

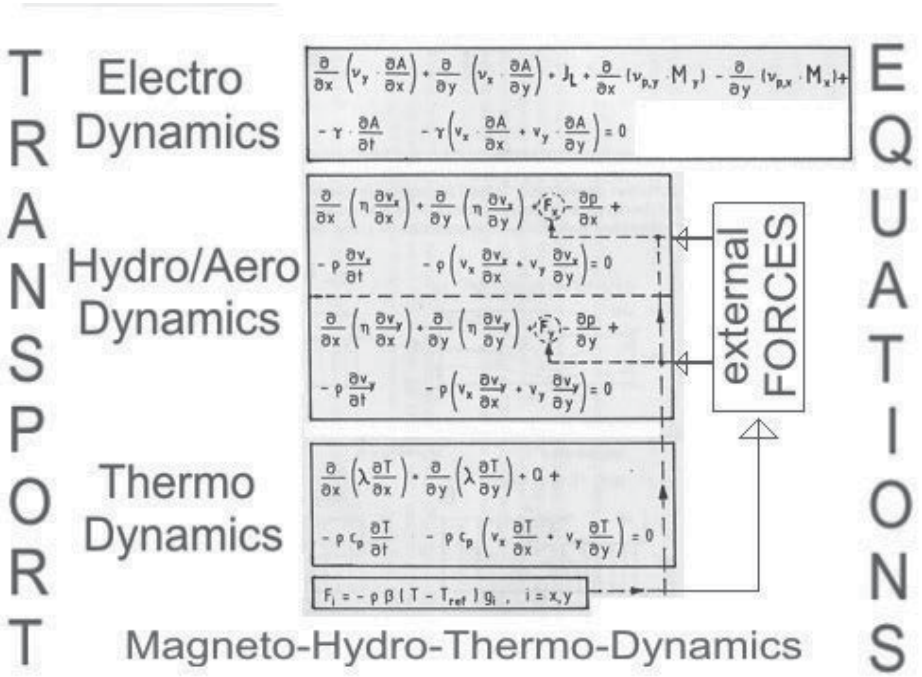


Fig. 7. Interdisciplinary structure identities of different and hybrid physical disciplines (Stanek, 2002).

Ferrohydrodynamic Bernoulli-Rosensweig equation based on magnetic fluids

The Bernoulli equation can be deduced from Navier-Stokes equations and extended with a magnetic polarisation $\mathbf{M}_p$ to handle i.e. industrial separation of diamonds:

$$p + \rho_m \cdot \mathbf{v}^2 / 2 + \rho_m \cdot g \cdot h + \rho_m \cdot \int (\partial \mathbf{v} / \partial t) d\mathbf{l} - \int \mathbf{M}_p \cdot d\mathbf{H} = const \quad (21)$$

Eq. (21) describes i.e. lifting of "stones" in magnetic fluids by magnetic fields (Stanek, 2002).

Magnetostriction and Electrostriction: Elastomechanics + Electrodynamics + Entropy

The deformation force densities $\mathbf{f}_{MS}$ (or $\mathbf{f}_{ES}$) of ferromagnetic (or dielectric) materials with density $\tau=1/\nu$ caused by magnetic (or electric) fields **H** (or **E**) can be derived by means of entropy. The converse effect applying mechanical pressure p to certain non-conducting crystals producing electric charges is piezoelectricity. All effects may depend on temperature T, too. In Fig. 8 interdisciplinary derivations are shown.

Maxwell → Poynting **E** x **H** → energy thermodynamics → "unavailable for work" entropy dS = dQ/T					(22)
Internal system energy dW_i = f [specific volume v(p), T, H (or E)] = transported heat dQ + total work dW_w					(23)
entropy dS_{MS} with dW_i=f(ν,H,T) for magnetostriction	$dW_{w,MS} + p \cdot dV = H \cdot d(\mu \cdot H) = \mu \cdot H\, dH + H^2 \cdot [(\partial\mu/\partial\nu)\, d\nu + (\partial\mu/\partial T)\, dT]$				(24)
	$dS_{MS} = dQ/T = \frac{1}{T}(\frac{\partial W_i}{\partial \nu} + \frac{p}{\nu} - H^2 \frac{\partial \mu}{\partial \nu}) d\nu + \frac{1}{T}(\frac{\partial W_i}{\partial H} - \mu \cdot H) dH + \frac{1}{T}(\frac{\partial W_i}{\partial T} - H^2 \frac{\partial \mu}{\partial T}) dT$..				(25)
	from eq. (22) - (25) → force density f_{MS} = - grad p(ν,H,T), neglecting T : eq. (26) - (27)				
Magnetostriction force density	$f_{MS} = \frac{1}{2} grad\left(H^2 \cdot \tau \cdot \frac{\partial \mu}{\partial \tau}\right)$	(26)	Electrostriction force density	$f_{ES} = \frac{1}{2} grad\left(E^2 \cdot \tau \cdot \frac{\partial \varepsilon}{\partial \tau}\right)$	(27)

Fig 8. Interdisciplinary entropy equations for magnetostriction + electrostriction (Simonyi, 1993; Stanek, 2002).

3. Interdisciplinary industrial applications in mechatronics

Four developments in the huge field of motor car production, magnetic gripper design in robotics, motor car anti-vibration systems and computer hard disk drives will demonstrate actual industrial applications in mechatronics based on electrodynamics.

3.1 Motor car production based on electrodynamics

In Fig. 9 we see a graphical overview of the actual topics of this publication concerning motor car production and influences. The field numerical optimisation of the actual holding and stacking system in world wide motor car production is the special focus in chapter 3.1.

The principle of mechanical motor car production is shown in **Fig. 9a.**

This computer-aided development was a combination of designing the necessary electrodynamic actuator (in **Fig. 9b**) and optimising both the aerodynamic flight of metal plates and the elastodynamics respectively plastic stacking in the stopping equipments.

This holding and stacking system, based on an international patent by Thyssen in Dortmund (inventors W. Stanek et al), has been used world wide in motor car plants for years (Stanek et al., 1984)

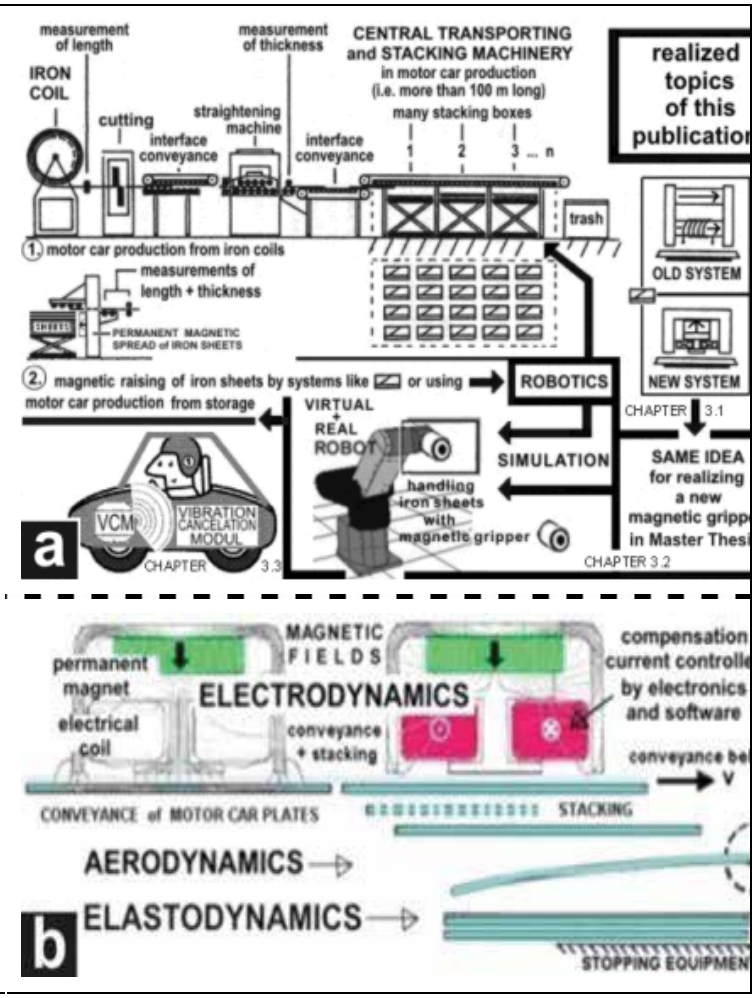


Fig. 9. Mechanical motor car production based on electrodynamics (Stanek, 2002+2010).

3.2 Gripper design in robotics based on electrodynamics

Based on the idea for the controlled actuator in motor car production, the following magnetic gripper was developed, simulated and realised in Fig. 10.

The 4 steps in the pictures shown on the right are necessary for each development with respect to electrodynamic actuators and sensors in robotics from idea to end product. This new magnetic gripper was developed by a Master's Thesis in Mechatronics in cooperation between SGU advisor W. Stanek from University of Applied Sciences Koblenz and Swiss German University (SGU) in Indonesia.

This development also shows the necessity for a mechatronics engineer to be flexible in working within several physical disciplines, including automated production, complex environments and both software controls and simulations.

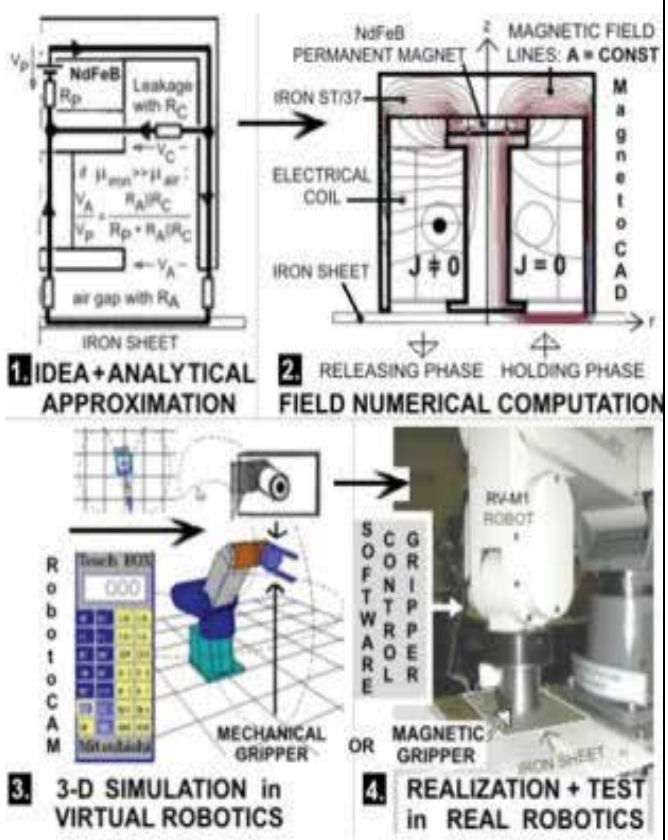


Fig. 10. New magnetic gripper design for handling metal sheets in 4 steps (Andries, 2003; Stanek & Grueneberg, 2003).

3.3 Motor car anti-vibration system based on electrodynamics

The cancellation of noise inside motor cars, using software controlled actuators, is of great importance in all motor car plants. The design of such anti-vibration systems (i.e. VCM) in Fig. 11 involves several interdisciplinary areas in physics such as acoustics, electrodynamics, thermodynamics, hydrodynamics, mechanics, elastodynamics and sound design, too.

Integrated simulation processes were needed for mechanical design of spring and damping in relation to electrodynamically and mechanically produced forces.

This new actuator, integrated in the software controlled VCM-circuit was parametrised, optimised and realised by a Diploma Thesis at Trelleborg Automotive in cooperation with Prof. Stanek (University of Applied Sciences Koblenz).
In the pictures shown on the right the focus on electrodynamics is emphasised.

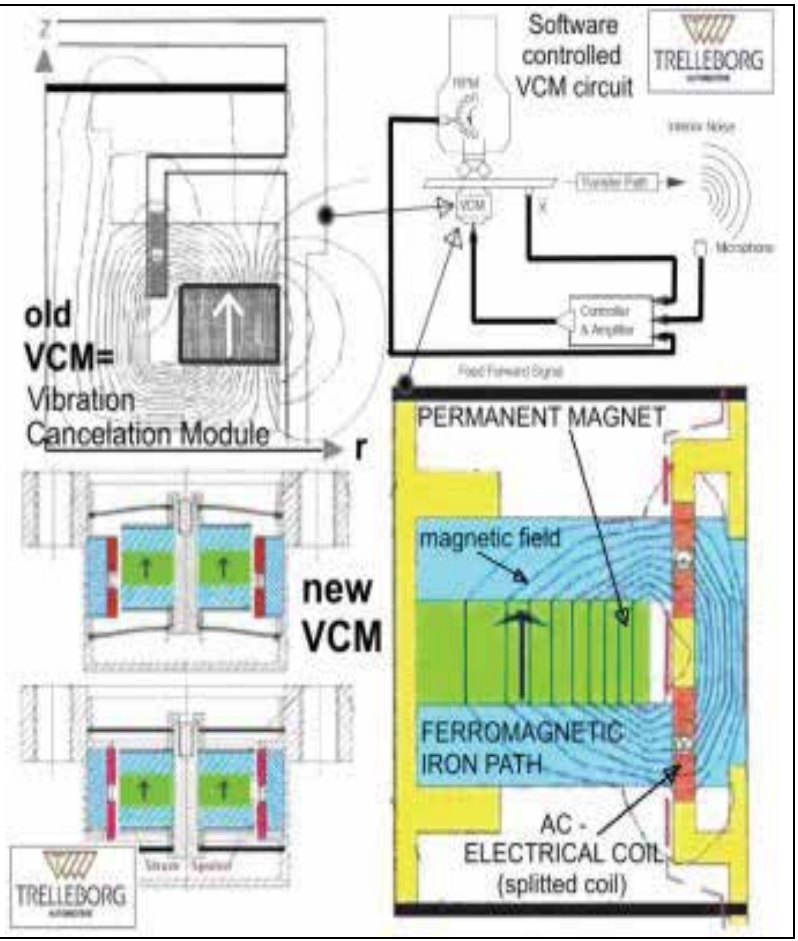


Fig. 11. Motor car anti-vibration system based on electrodynamics R&D report 2000 (Stanek, Graeve, Loehr, 2001)

3.4 Computer hard disk drives

3.4.1 Basic construction and basic governing equations

Complex concepts and applications of electrodynamics are the basis for a great variety of Hard Disk Drives in computers (i.e. often like in Fig. 12 or other special variants). Though very different in construction details, all hard disk drives are consisting of electrical coils, permanent magnets, iron parts and often additional copper plates or closed coils in form of a "shorted turn" (ref. to Fig. 13 and 14, the principle of a Winchester-Hard-Disk-Drive). The main task of these drives is to perform and to control the accelerated movements of magnetic heads for an exact and fast reading and writing of data on the magnetic hard disk.

Combined analytic modelling and computer aided simulation of mechanical and electromagnetic devices in mechatronics is necessary to solve and to simulate the behaviour of computer hard disk drives, i.e. Winchester drives.

For analytical calculation of electromagnetic fields in mechatronic systems and interdisciplinary analogies: Thinking in magnetics with concentrated field elements (R, L and often C) and solving dynamics by well known electrical circuit methods (i.e. with MATLAB, Simulink aided by FEMLAB, MAXWELL & MagnetoCAD) Simulation, Design in cooperation between W. Stanek and SGU Mechatronics

Fig 12. Often constructed hard disk drive (photo)

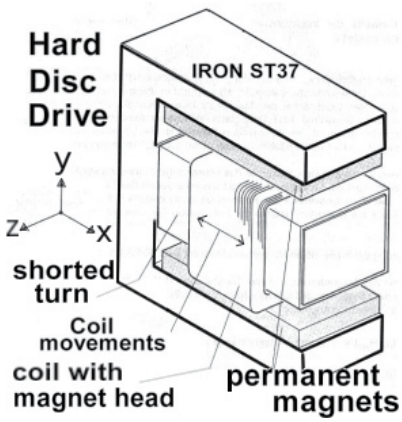


Fig 13. Principle of Winchester hard disk drive (Cassing, Stanek, Erd, 2002)

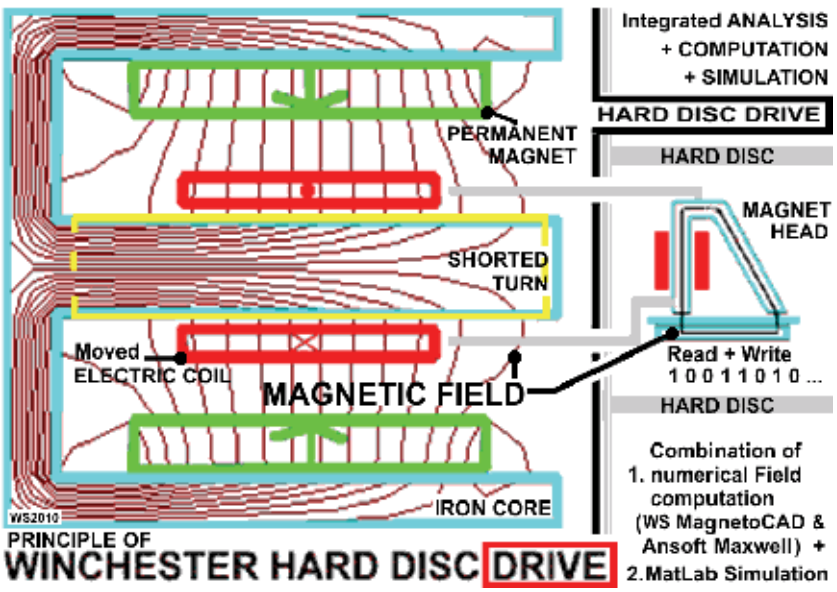


Fig 14. Winchester hard disk drive with magnetic field values for eq. (28-33) and eddy current equation from Maxwell (Stanek, 2008)

3.4.2 Modeling, analysis and simulation of Winchester hard disk drive unit

As one application for the significance of unit checks and its application to mechatronical system modeling and design, mechanical system of a computer hard disk drive is being explored. The physical structure may be seen in Figure 12 & 13 while the pattern of magnetic field inside the disk drive may be seen in the Figure 14.

The analytical equations (28 – 31) are shown later considering **N**o "**S**horted turn" (index "**nS**"). Modeling the influence of the "**S**horted turn" (index "**S**") as a transformer with "1" turn on the secondary side we can use the equations (32 & 33). The flow pattern of the magnetic field in a Winchester hard disk drive can be seen from the Figure 14, where there are two windings, the moving electric coil and the shorted turn. The corresponding magnet head will directly move with the movement of the electric coil.

We will analyse first the relationship of each parameter of this electromechanical system, as to produce the inter-connected equations needed to build a transfer function which express the output as the function of input parameter. This is started with the unit checks, and will be explained as far as the design and performance of the system response.

From the electrical circuit law, the current conducting coil will give relationship:

$$u_{1,ns}(t) = R_1 i_1(t) + \frac{d\left(L_1 i_1(t)\right)}{dt} + u_{ind}(t) \tag{28}$$

- A particle or body moving with a uniform speed *v* in a magnetic field *B:*

  Then by analysing the units check for $\vec{v} \times \vec{B}$ is $\frac{m}{s} \bullet \frac{V.s}{m^2} = \frac{V}{m}$ which shows that is the unit of Electric field, $\vec{E}$, that confirms the Lorentz Law of Electric Field, $\vec{v} \times \vec{B} = \vec{E}$.
- A current carrying conductor moving with speed v in **B** is given by $\vec{J} \times \vec{B}$, with unit check gives $\frac{A}{m^2} \bullet \frac{Vs}{m^2} = \frac{Ws}{m^4} = \frac{Nm}{m^4} = \frac{N}{m^3}$. And the result confirms the relationship of Lorentz law of spatial force density, $\vec{J} \times \vec{B} = \vec{f}$.
- The relationship of induced voltage which is the closed line integral of the electric field is given by $u_{ind}(t) = \oint \vec{E} dl$ and as l is constant then $u_{ind} \propto \vec{E}$. Given $\vec{E} = \vec{v} \times \vec{B}$ then as $\vec{B}$ is perpendicular to $\vec{v}$, then it can be represented as a scalar product $\vec{E} = \vec{v} \bullet \vec{B}$. As B is constant then it may be stated as $\vec{E} \propto \vec{v}$. Because of perpendicular relation of r & ω then $\vec{v} = \vec{r} \times \vec{\omega}$ can be reduced to, $\vec{v} = \vec{r} \bullet \vec{\omega}$,therefore with r being constant $v \propto \omega$, which yields:

$$u_{ind}(t) = k_1 . \omega(t) \tag{29}$$

- The torque equation may be expressed as:

$$\frac{d\left(J_{mec} . \omega(t)\right)}{dt} = T_m(t) - T_L(t) \tag{30}$$

For the moving Torque $\vec{T}_m = \vec{r} \times \vec{F}$ as perpendicular to each other then $\vec{T}_m = \vec{r} \bullet \vec{F}$ so $\vec{T}_m \propto \vec{F} \propto \vec{f}$. And as $\vec{f} = \vec{J} \times \vec{B} \propto \vec{J}$ due to B is constant, the result is $f \propto i$ and we get:

$$u_{ind}(t) = k_1.\omega(t) \tag{31}$$

Similarly for the condition of considering the influence of the "shorted turn":
The electrical circuit equations may be expressed as:

$$u_{2,s}(t) = R_1 i_1(t) + \frac{d(L_1 i_1(t))}{dt} + u_{ind}(t) + \frac{d(Mi_2(t))}{dt} \tag{32}$$

and

$$0 = R_2 i_2(t) + \frac{d(L_2 i_2(t))}{dt} + \frac{d(Mi_1(t))}{dt} \tag{33}$$

If inductances L_1, L_2, mutual inductance M and mass moment of inertia J_{mec} are constant, equations (28-33) can easily be simplified: i.e. $d(L \cdot i)/dt = L\, di/dt + i\, dL/dt$, where the last term is zero. MATLAB® and Simulink are mighty systems for simulating problems in mechatronics. But without the numerical computation of central electromagnetic field values, primarily L and M, analytical simulations may yield false results not usable for optimised applications in practice. Directly from equations (28-33) we can sketch the block diagram and the automation graph.

Design considering no shorted turn

	Time Relation	Laplace Transformation
The Motor Torque	$T_m(t) = K_2 i_1(t)$	$T_m(s) = K_2 I_1(s)$
The Voltage induced	$u_{ind} = K_1 \omega(t)$	$U_{ind}(s) = K_1 \Omega(s)$
Electrical Circuit Eq.	$u_{1,ns}(t) = R_1 i_1(t) + \frac{d(L_1 i_1(t))}{dt} + u_{ind}(t)$	$U_{1,ns}(s) = (sL_1 + R_1) I_1(s) + U_{ind}(s)$
Torque Relations	$\frac{d(J_{mec}.\omega(t))}{dt} = T_m(t) - T_L(t)$	$J_{mec} s\Omega(s) = T_m(s) - T_L(s)$

Rearranging all the Laplace form equation in the table and simplifying, the overall equations can be represented in the well-known block diagram relationship as shown in Figure 15.

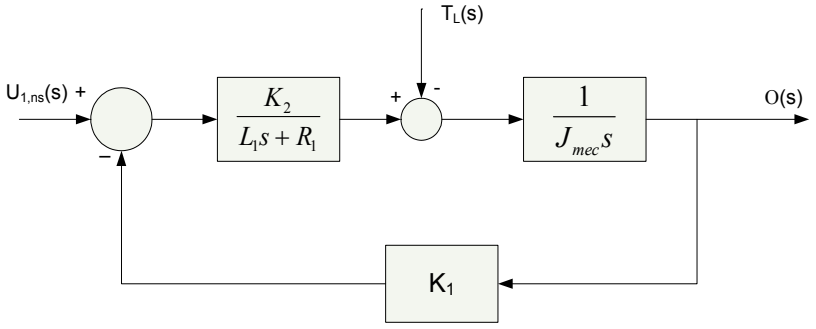


Fig. 15. Non-shorted hard disk drive System block diagram

By inserting general values of the variables and simulate the system response when subjected to a step input, we can obtain the system response of the angular velocity, current as well as torque produces shown on the next figures. The values chosen are general approximations based on common application of DC electric motor or hard disk drives, etc.
From the graph on the left, it can be seen the response of the angular velocity of the hard disk drive magnetic arm.
It can be seen that by the help of the MATLAB® and Simulink tools, the response of the model that has been designed may be observed. Graphical presentation of the current and the corresponding torque are shown below.

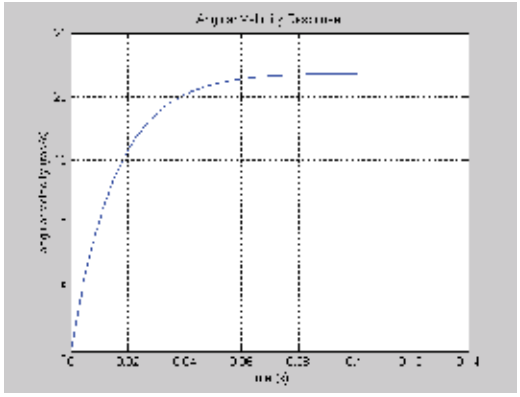


Fig. 16. a Angular Speed Response

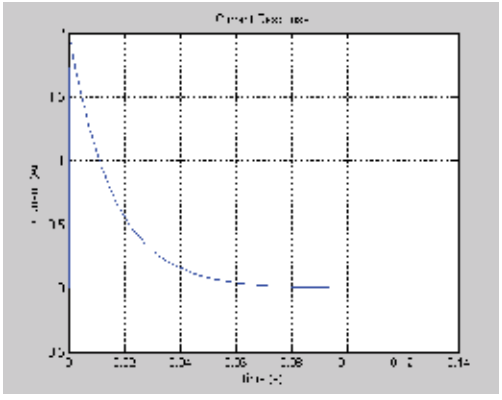


Fig. 16b Current Response

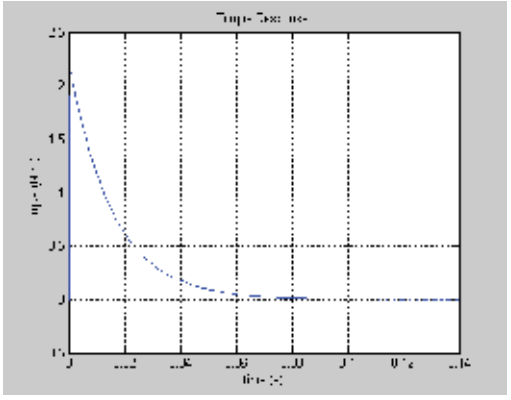


Fig. 16c. Torque Response

Fig. 16. Some excerpts of simulation results for a hard disk without shorted turn.

Also, it can be observed that the system response as depicted by Figure 16 resembles that of the system response of permanent magnet DC motor.

Design considering shorted turn

	Time Relation	Laplace Transformation
The Motor Torque	$T_m(t) = K_2 i_1(t)$	$T_m(s) = K_2 I_1(s)$
The Voltage induced	$u_{ind} = K_1 \omega(t)$	$U_{ind}(s) = K_1 \Omega(s)$
Electrical Circuit Equation (moving coil)	$u_{1,s}(t) = u_{1,ns}(t) + \frac{d(M.i_1(t))}{dt}$	$U_{1,s}(s) = (L_1 s + R_1) I_1(s) + U_{ind}(s) + MsI_2(s)$
Electrical Circuit Eq. (single turn)	$0 = R_2 i_2(t) + \frac{d(L_2 i_2(t))}{dt} + \frac{d(Mi_1(t))}{dt}$	$0 = (L_2 s + R_2) I_2(s) + MsI_1(s)$
Torque Relations	$\frac{d(J_{mec}.\omega(t))}{dt} = T_m(t) - T_L(t)$	$J_{mec} s\Omega(s) = T_m(s) - T_L(s)$

Rearranging all the Laplace form equations in the table and simplifying, the overall equations can be represented in the well-known block diagram relationship as shown in Figure 17.

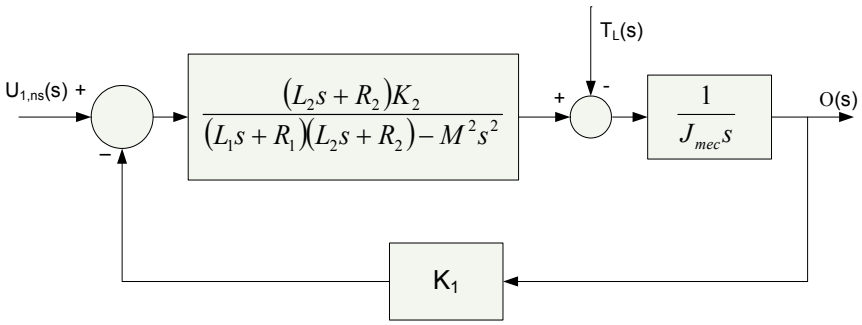


Fig. 17. Shorted hard disk drive system block diagram

From the transfer function in the block diagram, the order has increased by one and by analysing the characteristics equation in the first block diagram, and doing the similar methods as the non-shorted simulation shown previously which will give us necessary information or results that are required.

Mathematical and graphics tools such as MATLAB® & Simulink, are great tools to solve and describe the performance of the system, but it is most important to know that for engineering application, the understanding of electrodynamics is the key to obtain the model to be simulated by those tools.

The application of the magic unit checks for physics and extended field theory based on interdisciplinary electrodynamics in this Mechatronical System is successfully derived therefore it is possible to apply it on other mechatronic and automation systems.

4. Conclusion

The high aim of optimising the integration of mechanical engineering, electrical engineering and information technology in mechatronics can often be reached by preferred usage of advanced field theory in electrodynamics. Working with extended Maxwell's equations, electrodynamics in mechatronics often leads to new developments and interdisciplinary influences which are easier and faster to approximate. Quick derivation of interdisciplinary and complicated equations in physics can be achieved by using extremely helpful and mighty unit checks. Furthermore other electrodynamic influences especially caused by external waves and fields with respect to electro-magnetic compatibility problems can be handled and corrected.

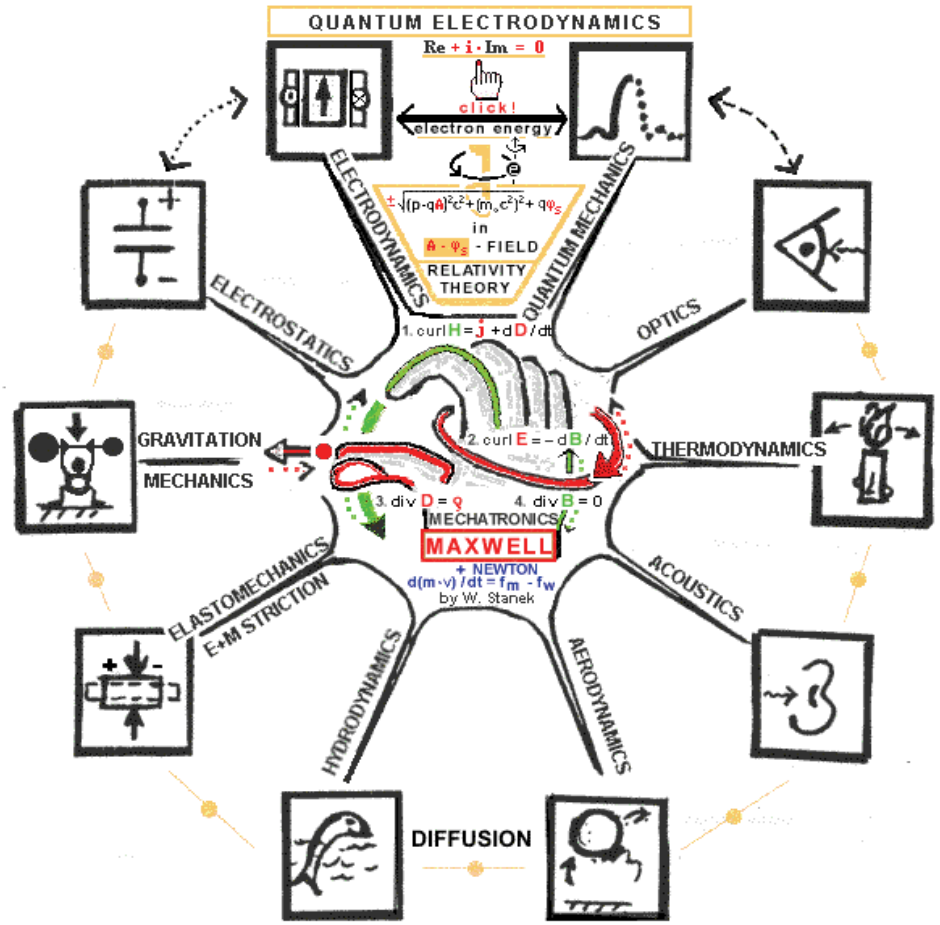


Fig 18. Interdisciplinary analogies in mechatronics based on extended Maxwell's equations (Stanek, 2002+ 2010)

Fig. 19. Compact equation for most central equations in classic, relativistic and quantum electrodynamics (Stanek, 2010)

The focus on four described developments such as motor car production, magnetic gripper design in robotics, motor car anti-vibration systems and computer hard disc drives show the necessity for a mechatronics engineer to be flexible in working with several physical disciplines (refer to Mind Map with Memo Maps in Fig. 18), in highly automated car production including complex environments and with both a variety of different software controls and simulations. Most central equations for the interdisciplinary engineers and physicists can be derived from the very compact equations shown in Fig.19. This unified equation for relativistic quantum electrodynamics applies most central relations and analogies from section 2 and spectrum of interdisciplinary disciplines in advanced mechatronics. It is important to use a spectrum of background disciplines rather than just one discipline.
The compact equation for most central equations in extended electrodynamics was developed by the main author W. Stanek based on Faraday's law, Einstein's relativistic energy, including quantum mechanics in complex notation.
This compact unified equation, Re + i . Im = 0, consists of Maxwell's equations in rest and moving bodies, Lorentz-Einsteins' relativistic energy relations, Klein-Gordon's equations, relativistic Schroedinger's equation, Proca's extended Maxwell's equations, central relations in quantum mechanics and classical Newton mechanics itself, too.

5. Acknowledgements

Thanks of main author Prof. Dr. W. Stanek to his appreciated colleagues and co-authors Ir. Arko Djajadi, Ph.D and Pro Rector Edward Boris P Manurung, MEng from SGU – Asia for adaptation and MATLAB® & Simulink simulation based on W. Staneks' research and publication (refer to list of publications).

6. References

Cassing, W., Stanek, W. et.al.: „Elektromagnetische Wandler und Sensoren",
Chap.1: Electrodynamic, computer-aided development of actuators and sensors
Chap.2 - 8: Applikationen aus allen Bereichen der Technik, Expert-Verlag, Renningen, 2002.

Andries, R.: Design, Field Simulation and Realisation of a New Magnetic Gripper for Handling Metal Objects in Real and Virtual Robotics, Master Thesis, Advisor W. Stanek, both at Swiss German University (SGU) BSD-Jakarta, Java, and FH Koblenz, Germany 2003.

Stanek, W., Greaeve, A., Loehr, B.: Design, Parametrisierung und Realisierung eines mechatronischen Schwingungssystems, Report Research and Development 2000 FH Koblenz and Trelleborg Automotive, WEKA-Verlagsgesellschaft, Koblenz, 2001.

Bronstein, I. N. et.al.: Teubner-Taschenbuch Mathematik Teil I+II, Teubner-Verlag Stuttgart, 1995.

Lehner, G.: Elektromagnetische Feldtheorie für Ingenieure und Physiker, Springer-Verl., Berlin,1994

Simonyi, K.: Theoretische Elektrotechnik, Barth Verlag, Bad Langensalza, 1993

Sommerfeld, A.: Elektrodynamik, Verlag Harri Deutsch, Thun, 1988

Stanek, W. et.al. a) „Gedächtnistraining – Das Erfolgsprogramm für Neues Lernen" Goldmann-Verlag, München, 2006 + b) Ripol Publishing House, Moscow, 2009 (upgrade in Russian)

Stanek, W.; Huebner, K.D.; Oettinghaus, D.: „Permanent magnetic charge taking or holding device", Patent, Publication-Numbers: EP000000182961A1, US000004594568A, Thyssen Germany, 1984

Stanek, W.: Extended Maxwell's Equations: compact formulations in physics, http://www.wolfram-stanek.de/maxwell_equations.htm, including detailed furtherlinks, 2010

Stanek, W.; Grüneberg,J.: Electrodynamics and its analogies in physics … , REM conference on Research and Education in Mechatronics, Bochum, 2003, Common publication of FH Koblenz, Germany, & Swiss German University – Asia (SGU), Indonesia

Einstein, A.: Zur Elektrodynamik bewegter Körper, Annalen der Physik und Chemie, Jg.17, Bern, 1905
http://de.wikibooks.org/wiki/A.Einstein_Zur_Elektrodynamik_bewegter_Körper._Kommentiert_und_erläutert

Stanek, W.: Lectures on "Mechatronics" at intl. universities, i.e. SWISS GERMAN UNIVERSITY-ASIA, Java-BSD, 2002, 2003, 2010; Catholic University Indonesia ATMA JAYA, Jakarta, 2003; Technical University Opole, Poland, 2005-2008.

Oberretl, K.: Appreciated review of the main author's book "Elektromagnetische Wandler und Sensoren" by Univ.-Prof. Dr. Kurt Oberretl, University Dortmund, 2008

Edited by Igor Fuerstner

Today's global economy offers more opportunities, but is also more complex and competitive than ever before. This fact leads to a wide range of research activity in different fields of interest, especially in the so-called high-tech sectors. This book is a result of widespread research and development activity from many researchers worldwide, covering the aspects of development activities in general, as well as various aspects of the practical application of knowledge.

ISBN 978-953-51-5144-9

Photo by Tzido / iStock

IntechOpen